DOCUMENT RESUME

ED 148 329 IR 005 372

AUTHOR Tallmadge, G. Kasten
TITLE The Joint Dissemination Review Panel IDEABOOK.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Oct 77
CONTRACT NIE-IA-7706
NOTE 101p.

EDRS PRICE MF-$0.83 HC-$6.01 Plus Postage.
DESCRIPTORS *Adoption (Ideas); Comparative Analysis; Data Collection; Diffusion; *Educational Assessment; *Educational Equipment; *Educational Innovation; *Educational Programs; *Evaluation Methods; Federal Programs; Information Dissemination; Instructional Innovation; School Superintendents
IDENTIFIERS Joint Dissemination Review Panel

ABSTRACT

The Joint Dissemination Review Panel (JDRP) is a joint USOE-NIE panel which reviews the evidence of effectiveness submitted for a variety of educational products and practices developed through federally supported programs. Studies show that school superintendents and other adopters prefer to initiate changes where there is assurance that benefits are reliable, valid, and substantial. JDRP provides this assurance. This Ideabook, prepared at the request of practitioners, has two purposes. The first is to illustrate the ways to gather convincing evidence of the effectiveness of educational innovations. It is, thus, a guide for planning evaluations in anticipation of future submission to the JDRP. The second purpose is to suggest how to bring current evidence of effectiveness together in a succinct and forceful manner. Such presentations enhance the chances for JDRP approval and enable potential adopters to compare products and their associated claims. (Author/DAG)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED148329

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Joint Dissemination Review Panel

IDEABOOK

By

G. Kasten Tallmadge

RMC Research Corporation
Mountain View, California

October, 1977

U. S. Department of Health, Education and Welfare
Joseph A. Califano, Jr., Secretary
Mary F. Berry, Assistant Secretary for Education
NATIONAL INSTITUTE OF EDUCATION
Patricia Albjerg Graham, Director
U. S. OFFICE OF EDUCATION
Ernest L. Boyer, Commissioner of Education

Washington, D. C.

IR005372

This report was funded by the National Institute of Education under Contract Number NIE-IA-7706. Its contents do not necessarily reflect the views of the Institute or of any other agency of the United States Government.

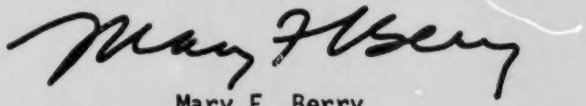
FOREWORD

During the four years of Joint Dissemination Review Panel operation over three hundred submissions have been reviewed. Studies show that school superintendents and other adopters are strongly influenced by JDRP approval. They prefer to initiate changes for which there is reasonably strong assurance that the benefits claimed for students are reliable, valid, and substantial. The JDRP's review of evidence of effectiveness provides this assurance.

The past four years have also seen a substantial improvement in the quality of submissions to the JDRP. The quality of the information collected about the effectiveness of new developments in education appears to have increased, and much of this improvement can be attributed to the JDRP's review procedures. With the increasing number of submissions this influence will probably grow.

This Ideabook, prepared at the request of practitioners, has two purposes. The first is to illustrate the many ways to gather convincing evidence of the effectiveness of educational innovations. It is, thus, a guide for planning evaluations in anticipation of future submission to the JDRP. The second purpose is to suggest how to bring current evidence of effectiveness together in a succinct and forceful manner. Such presentations not only enhance the chances for Panel approval, but also help potential adopters better to compare products and their associated claims.

While the Ideabook and the Joint Dissemination Review Panel itself are only two of several approaches developed by the Education Division to get products that work into the hands of practitioners, we believe the JDRP Ideabook will be particularly useful.



Mary F. Berry
Assistant Secretary
for Education

ACKNOWLEDGEMENTS

Many people made significant contributions to this document. The author is most indebted to Lois-ellin Datta, current Chairperson of the JDRP and Assistant Director of NIE's Education and Work Group. Dr. Datta provided great assistance in preparing the outline, reviewing and editing the draft report, and preparing the illustrative examples that appear throughout the text. Other reviewers include:

Janice K. Anderson, Office of Planning, Budgeting, and Evaluation, U. S. Office of Education

Naida T. Bagenstos, Research and Development Interpretation Service, CEMREL, Inc.

Glen C. Boerrigter, Research Branch, U. S. Office of Education

Albert C. Crambert, Planning and Special Programs, Region III, U. S. Department of Health, Education, and Welfare

Mary Kennedy, Bureau of Education for the Handicapped, U. S. Office of Education

Susan S. Klein, Dissemination and Resources Group, National Institute of Education

Diane McIntyre, Far West Laboratory for Education Research and Development

Their comments were most helpful in preparing the current version of this JDRP Ideabook. Constraints of time and space prevented the inclusion of some of their recommendations, but most I have gratefully adopted. Despite my heavy reliance on help from others, the responsibility for whatever shortcomings remain is mine alone.

G. Kasten Tallmadge

CONTENTS

FOREWORD, by Mary F. Berry, Assistant Secretary for Education i

ACKNOWLEDGEMENTS iii

I. WHAT IS THE JOINT DISSEMINATION REVIEW PANEL AND WHY DOES IT EXIST? 1

II. WHAT ARE THE ESSENTIALS OF AN ADEQUATE EVALUATION?. 5

III. CRITERION 1: DID A CHANGE OCCUR?. 9

IV. CRITERION 2: WAS THE EFFECT CONSISTENT ENOUGH AND OBSERVED OFTEN ENOUGH TO BE STATISTICALLY SIGNIFICANT? 25

V. CRITERION 3: WAS THE EFFECT EDUCATIONALLY SIGNIFICANT? 31

VI. CRITERION 4: CAN THE INTERVENTION BE IMPLEMENTED IN ANOTHER LOCATION WITH A REASONABLE EXPECTATION OF COMPARABLE IMPACT? 41

VII. CRITERION 5: HOW LIKELY IS IT THAT THE OBSERVED EFFECTS RESULTED FROM THE INTERVENTION 47

VIII. CRITERION 6: IS THE PRESENTED EVIDENCE BELIEVABLE AND INTERPRETABLE?. 53

IX. COMMON EVALUATION HAZARDS AND HOW TO AVOID THEM 61

X. FORMAT AND INSTRUCTIONS FOR SUBMITTING MATERIALS TO THE JOINT DISSEMINATION REVIEW PANEL 73

XI. A MODEL SUBMISSION 77

XII. ANOTHER MODEL SUBMISSION 85

REFERENCES 97

SOME SOCIAL FAIRNESS GUIDELINES. 99

I. WHAT IS THE JOINT DISSEMINATION REVIEW PANEL AND WHY DOES IT EXIST?

The Joint Dissemination Review Panel (JDRP) was established by the Education Division of the Department of Health, Education, and Welfare (HEW) in 1972. Originally a part of the U.S. Office of Education (USOE), the review panel was enlarged in 1975 by the addition of representatives from the National Institute of Education (NIE). The joint USOE-NIE panel currently has 22 members, eleven from each agency, appointed respectively by the Commissioner of Education and the Director of NIE. The members are chosen for their experience in education and their ability to analyze evaluative evidence on the effectiveness of educational products and practices.

The JDRP meets periodically to review the evidence of effectiveness submitted for a wide variety of educational products and practices. While the products and practices must have been developed through Federally supported programs, they may be copyrighted or published commercially. Submissions are made directly to the panel by USOE and NIE program offices, which are responsible for identifying potentially exemplary projects and for carrying out a pre-review screening for factual accuracy, social fairness, and possible harm to users, as well as for evidence of effectiveness. Only those submissions that are approved by the JDRP may be in any way endorsed by the Education Division or disseminated as exemplary using Federal Education Division funds.

The full review process begins with an Education Division project officer who believes that an intervention supported by Education Division funds may be worthy of dissemination as exemplary. Different program offices within the Education Division employ different strategies for initially identifying promising interventions. Once identified, however, all interventions must be reviewed by the program office for social fairness of content and procedures, and for honesty, completeness, and accuracy of the evidence supporting the intervention's effectiveness. While the results of these pre-submission reviews are not forwarded to JDRP, they are available at the submitting program office.

Submission to the JDRP is by means of an official transmittal memorandum that must be signed either by an NIE

Associate Director or a USOE Assistant Commissioner. The submissions are received by the Panel's Executive Secretary, who arranges the review as soon as three submissions have accumulated.

Reviews are conducted by nine-member sub-panels of the JDRP convened by the Executive Secretary. At least four members of each review body represent NIE and at least four represent USOE.

The submitting project officer is notified of both the date of the review and the membership of the reviewing sub-panel as soon as arrangements are made. The meetings are open to the public and detailed minutes, also a matter of public record, are kept. Votes are taken on each submission immediately after review and interventions are "passed" by a simple majority. Where the presented evidence is less than adequate, the JDRP may suggest resubmission.

Since its founding, the growth of the JDRP has been accompanied by an increasing recognition of the importance of its mission: To ensure that educational interventions--projects, products, or practices--have been shown to have positive impacts on their recipients before they are disseminated with Federal monies or endorsed by the Education Division.

Effectiveness is the sole criterion for approval by the JDRP. No judgments are made regarding the desirability of projects or products from any other standpoint. "Effectiveness," however, is not defined simplistically. In order for the JDRP to determine that a given product, project, or practice is effective, several conditions must be met. The evidence must be valid and reliable, the effect must be of sufficient magnitude to have educational importance, and it should be possible to reproduce both the intervention and its effects at other sites. These considerations give rise to the specific criteria used by the panel in judging the evidence of positive impact. (These criteria are discussed in detail in Chapters III-VIII.)

In keeping with the diversity of educational goals and innovations, it is recognized that there are several ways in which each of the criteria may be met. Before dealing with the criteria, however, it will be useful to consider

what a positive impact might look like "on the face of it" and what features might lead us, after further examination, to conclude that we are safe in taking this apparent impact at its face value.

There are many possible kinds of positive impact in educational programs. One example is gains: Students learn to read better, or they learn more in math. There are other possibilities: An intervention might result in improved attitudes toward school. Students might develop more positive self-concepts. They might have better mental or physical health. Positive impacts might affect others besides the students; an example would be improved instructional behaviors of teachers. Or reduction of sexual discrimination throughout a school might be considered a desirable result. Positive impacts are probably as numerous and as varied as the goals of our educational system.

Identification of an intervention with an apparent positive impact is the first step toward determining whether such an impact has actually occurred as the result of the intervention. In order to establish that a project is effective, there must be adequate evaluation data to tie the project and the desired impact together in a cause-and-effect relationship.

II. WHAT ARE THE ESSENTIALS OF AN ADEQUATE EVALUATION?

An adequate evaluation must be interpretable and credible. First of all, the existence of the apparent effect must be established. Then there must be evidence that the effect occurred as the result of the particular educational intervention, and that it would not have occurred without the intervention. These requirements, and ways of meeting them, will be discussed in detail in subsequent chapters. This chapter is intended to provide a perspective for that discussion.

Valid and Reliable Measures of Post-Intervention Conditions

Clearly, establishing the existence of an educational effect implies that there must be valid and reliable measures. A valid measure is one that has a logical relationship to what is being measured, whether it be reading, self-concept, or Board of Education decision making. Consider, for example, a project designed to enhance the decoding skills of low-achieving students. The children are pretested to measure their status before the project begins and posttested at the end. Where the instruments used are relevant to the content of the instruction, a valid assessment of progress can be made. If, however, the pupils were tested on vocabulary or comprehension, it is not likely that the obtained scores will reflect the outcomes of the instruction. If, on the other hand, we use a test measuring word attack skills, it is more likely to yield a valid assessment.

As another example, a project intended to increase parent involvement in the school might validly report increased parent visits, greater attendance at parent/teacher meetings, or greater knowledge of and participation in local school board elections. Reporting how well the parents like the program, would not, however, be compelling evidence of its effectiveness.

When the measures being used are directly related to the purpose and the content of an intervention, we can feel confident that we are measuring the right things, but it is also essential to know whether the measures are reliable. When a measure consistently ranks the same things in the same order, time after time, we say that it is reliable. If we measured the head sizes of all the students in a class and

found that each time they ranked in the same order from small to large, we would feel that our measuring instrument was a reliable one. In this situation, a measuring tape would be a reliable instrument; a rubber band would not.

A reliable instrument, unfortunately, is not necessarily a valid one. For example, if we measure the height of all the children with a measuring tape, we will get a reliable ranking from short to tall. But if we then say that this measure will be used to assess physical well being, we will not have a valid test.

What are the kinds of measures that may be used? As long as they are valid for the goals of the intervention, and are reliable in determining the size of effects, many types of measures are acceptable. There are measures that are obtained directly from the project participants which include:

- Achievement test scores
- Scores on attitude or self-concept scales

There are direct observational measures of changes the intervention was designed to achieve, for example:

- Measures of absenteeism
- Classroom observations of teaching behavior
- Measures of services delivered
- Indicators of parent involvement

Finally, there are unobtrusive measures designed to assess the achievement of objectives indirectly, by focusing on events that are signs of the desired outcome, not necessarily the outcome itself. Examples include:

- Measures of vandalism
- Number of books checked out of library
- Usage by women of facilities formerly restricted to men

Credible Estimates of Conditions Without the Intervention

In addition to valid and reliable measures of conditions after the intervention, an adequate evaluation requires credible estimates of what these conditions would have been

without the intervention. In other words, are things better than what we would have expected if there had been no special educational product or practice? This question implies a comparison--a weighing of the results of a particular intervention against those of no intervention, or of some other type of intervention. Such comparisons are critical, because in educational situations, we must always expect that some growth or change will occur even with no intervention. Some change will arise from learning experiences outside of the intervention or even simply from maturation. Since it is impossible to observe directly how a particular group of students would have performed without the intervention, it is necessary to rely on estimates of this no-treatment condition.

Again, these estimates can come from various sources. For example, although we cannot observe the same subjects both as participants and as non-participants at the same time, we might be able to observe a nearly identical group of non-participants in order to get an estimate of the no-treatment outcomes. Or we might get an estimate from the performance of participants prior to the intervention. The important thing is to find the most believable comparison--the one that will make us the most confident that the innovation was an improvement over what was done before. Examples of data that could be used for these comparisons include:

- Test scores of a control group
- National or local norms
- Historical records
- Amount of vandalism in neighboring schools
- Prior enrollment of limited-English students in vocational programs
- Former per-pupil cost of services

An educational intervention is like any other kind of experiment; we may hypothesize that it will prove to be an effective solution to a particular problem but, when we look at the results, we must be willing to consider other possible causes of the effects that were observed. In a laboratory this can be done by designing a carefully controlled experiment where all variables other than the one of interest are held constant. In educational situations this is simply not possible.

During the course of a project, many things, including the students themselves, may change as the result of events or influences that have nothing to do with the educational intervention. For this reason, when judging the effectiveness of any educational product or practice, the JDRP must determine not only that there was an effect, but that the effect most probably resulted from the intervention and not from some other cause. An appropriate comparison may provide a strong argument for the effectiveness of an intervention, but it is still necessary to consider the possibility that something else, acting at the same time as the intervention, was actually responsible. For example, if an evaluation compared the performance of all students in a district before and after a new reading project was installed, and if reading achievement had clearly improved, the panel might still want to inquire whether there could be some other explanation--an influx of higher achieving students into the district, or the introduction of Sesame Street by the local TV station.

As another example, a counseling program intended to prevent dropouts reported a substantial increase in high school graduation rates when compared to the rates in comparison schools. However, in the same period, the school had also expanded a work-study project designed to help students stay in school. When the dropout rates for students who participated in counseling only, in work-study only, and in both were compared with rates of eligible youth in the control schools, it was clear that it was the counseling program that was responsible for improved retention.

A well designed evaluation will always provide for an appropriate comparison and will attempt to control, or at least to account for, other possible causes of observed effects.

III. CRITERION 1: DID A CHANGE OCCUR?

Before an intervention can be judged successful, there must be convincing evidence that some change occurred. The occurrence of change is thus a necessary, although not a sufficient, condition for effectiveness.

Actually this first criterion may be expanded somewhat to read as follows, "Was the change a positive one that was in some way related to the objectives?" A remedial reading project, for example, would not be considered by the JDRP unless it had benefitted the children served, even if it did have a positive impact on the community by providing employment for one teacher and three paraprofessional aides.

The requirement that outcomes be related to objectives is not intended to preclude acknowledgment of unexpected outcomes--as long as the outcomes can be logically related to the treatment. It may happen, for example, that a project designed to improve the self concepts of participating children will also result in the improvement of their achievement test scores. The fact that these outcomes were unexpected would not preclude JDRP approval.

What is a Change?

A change is an outcome or consequence claimed for the innovation that is logically related to it. In almost any educational project, numerous events might seem at first to be this kind of change: learner needs were surveyed, teacher competencies were assessed, workshops orienting parents to the program were held, students talked up more and acted out less in class, more adults signed up for the new program than could be served, visitors to the projects liked what they saw, newspapers wrote feature articles about the innovation, participants kept logs of their experiences. Some of these events are examples of an outcome change; some are not.

The first task is to identify for the JDRP those changes that seem to be:

- consequences of the innovation,
- central to the claims of effectiveness, and

- measured so as to provide credible evidence that the outcome objectives were achieved.

As an example, less than eleven percent of the migrant children who enter high school ever secure a high school diploma. One reason for this loss was found to be lack of exchanges for partial academic credit among states in the migrant stream. Teenagers who left a southwestern state after ten weeks in school are enrolled in the northern state schools with only four weeks left in the semester, would lose all academic credit. The innovative program involved recruiting, counseling, afternoon and evening schools, and credit exchange arrangements. The central evidence of change presented was, appropriately, school credit accumulation, school retention, and graduation. Evidence of academic progress as measured by achievement, while important, was considered as supporting, rather than central to the claims made. Evidence of school system, parent, and student attitudes was not considered central enough to present at length in the JDRP submission.

In contrast, a project whose stated goal was preparing handicapped children for regular primary school through a parent-centered preschool, offered as evidence of change parent satisfaction with the program, parent attitudes toward their children, fall-to-spring IQ gains, extensive teacher observations of changes in the children's social maturity during the preschool year, and distributions of how many program graduates were in regular classes without requiring special attention, in regular classes but requiring special attention, and in special education. The mean pre-treatment IQ scores were presented, but no other evidence was presented of pre-existing handicapping conditions (children were identified by parents as needing preschool help), and there was no evidence of how the post-program distributions would have been without the preschool. In this case, a great deal of marginally relevant information was presented while data directly related to the central claims were inadequate. Had more care been taken in organizing the submission, these deficiencies would have become apparent and could have been remedied.

Despite the need to select appropriate, central outcomes as evidence that something happened, claims can be as diverse as the range of educational objectives. As examples:

- dropout prevention can be measured by retention and subsequent completion

- helping migrant children keep pace with non-migrants can be measured by placement in regular classes, age-expected cumulation of academic credit, and performance on tests of skills and knowledge

- improved teacher skills can be measured by criterion-referenced competency and knowledge tests, classroom observations, and master teacher reports

- improved school productivity associated with a new management system can be assessed by showing that quality and outcomes remained the same or increased while per-pupil costs decreased

- an environmental education program can present evidence of greater knowledge of environmental law, of ways to protect the environment while it is used for recreational or economic purposes, and concern for the environment

- improved reading can be measured by performance on criterion-referenced tests, norm-referenced achievement tests, ability in the next grade to handle reading requirements, observations of reading fluency, and accuracy in performance tests.

What is central and appropriate evidence that changes occurred is determined, then, by the consequences the innovation is intended to achieve. Identifying this signal from the noise of all the events associated with the innovation is one purpose of the first criterion.

How Does One Demonstrate that a Change Occurred?

As pointed out earlier, there are two essential ingredients for a sound and objective evaluation. First, there must be some valid and reliable measure of conditions after the intervention has had time to "work." Second, there must be some credible estimate of what those conditions would have been, at the same point in time, without the intervention. It can be said that a change occurred if conditions after

the intervention are different from what they would have been without it.

There are many satisfactory ways in which these evaluation requirements can be met. Before discussing those strategies, however, it seems worthwhile to describe two widely employed approaches that are clearly unsatisfactory.

Unsatisfactory Approaches

The most prevalent example is the simple before-and-after comparison. While such comparisons are both valid and credible for displaying the advantages of mouthwashes, hair colorings, laundry detergents, and dancing lessons, they may be almost totally useless in other areas. The fact that 25 girls were observed to grow an average of 2.5 inches between their eleventh and twelfth birthdays, for example, would not necessarily prove that a special nutritional program in which they had participated was effective in promoting growth. Since the national norm for growth of eleven-year-old girls is approximately 2.9 inches, it might be inferred, in fact, that the special program had no benefit at all.

The same kind of situation exists within education. Simply showing that children score higher on a test after a year of instruction than they did before school began proves nothing. Since some growth is expected as a natural concomitant of simple maturation and normal out-of-school experiences, it is necessary to demonstrate that status after an intervention is higher than could be expected to result from such "regular" growth--not simply that post-intervention status was higher than pre-intervention status.

There is a second, closely related fallacy that applies specifically to standardized achievement tests, namely that if disadvantaged children make greater than month-for-month gains during the school year, they will move closer to the national norm. While this relationship must be true from a purely logical viewpoint, it is untrue in practice. The disparity arises from the manner in which grade-equivalent scores are derived by test publishers. On certain tests at certain grade levels, it is quite possible for disadvantaged children to make greater than month-for-month gains from fall to spring while falling farther behind the national norm.

This and other problems associated with norm-referenced evaluations are discussed further below. It is sufficient here to say that making month-for-month gains on standardized tests is not an adequate criterion of success for educational interventions serving below-average achievers.

How, then, can success be demonstrated? This question can best be answered in terms of the classical experimental model. Before proceeding, however, it must be pointed out that this model is presented primarily for illustrative purposes. It is not expected that it will be possible to implement the model under most real-world conditions. At the same time, however, it must be made clear that most other evaluation designs derive their credibility from the extent to which they provide reasonable substitutes for the elements of this classical experimental model.

The Classical Experimental Design

In the classical experimental design, two groups are formed either through random assignment or matching. Ideally the groups should be identical before the experiment begins. During the course of the experiment, the experiences of the two groups should be identical, with the single exception that one receives the experimental treatment while the other does not. Then, at the end of the experiment, some kind of status measures are collected. The scores of the treatment group provide a post-treatment status indicator while the scores of the no-treatment (control) group tell us what the scores of the treatment group would have been if the treatment had not been administered.

If all the conditions have been met, it is clear that we have the two essential ingredients of a sound evaluation: a valid and reliable measure of post-intervention conditions and a highly credible estimate of what those conditions would have been without the intervention.

The control-group model can also be implemented without random assignment of pupils to treatment and control groups. Typically, a non-participating classroom or school will be selected to serve as a control group because the students are similar to the students in the treatment group. Because pre-existing groups are almost certain to differ from one

another in systematic (even though small) ways, this design is usually called the nonequivalent control group design.

Quasi-Experimental Designs

The nonequivalent control group model is one example of a class of evaluation designs called quasi-experimental. Most of these designs are little more than variations on the control-group model, and it is quite helpful to think of them in that way. In general, the difference between quasi-experimental and true experimental designs lies in the way the estimate of without-intervention conditions is generated. When posttest scores of a randomly assigned control group are not available to provide this information, estimates can be derived from the following sources (listed in approximately descending order of credibility):

a. a highly similar but nonequivalent control group

b. historical (pre-intervention) data

c. a comparison group formed by dichotomizing an originally intact group into treatment and comparison components around some pretest cutoff score (Regression-discontinuity and regression-projection models. See Campbell & Stanley, 1963; Horst, Tallmadge, & Wood, 1975.)

d. national, regional, or local normative data

e. logic

f. expert testimony

It is beyond the scope of this paper to describe in detail how sound evaluations can be implemented using each of these estimation sources. For further discussion of evaluation designs, see Campbell and Stanley (1963), Horst, Tallmadge, and Wood (1975), and Tallmadge and Horst (1976). Some guidelines can be offered here, however, with respect to a few of the major issues. Perhaps the most important are those relating to norm-referenced evaluations since they, probably because of their feasibility of implementation, are the most commonly encountered.

The special case of norm-referenced evaluations. Sources a, b, c, and d all require some sort of empirical data. Consider, for example, the use of normative data. What we do in essence is compare the test scores of children in the treatment group against those of children in the normative sample to determine before-intervention and after-intervention status. If we find that our group had a higher percentile status after the intervention than before, we would be likely to conclude that the intervention was successful. Quite possibly, however, this inference could be distorted and it is useful to analyze exactly the rationale that underlies it.

Suppose our intervention was a sixth-grade remedial math project. Before it began, we administered a standardized math test to the children who were to be served, calculated a mean score for the group, and looked it up in the test publisher's beginning-of-sixth-grade national norms table. We found that the mean score for our group fell at the 20th percentile. At the end of the school year we followed the same procedure and found that our group mean score fell at the 30th percentile of end-of-sixth-grade children.

When we infer from these findings that our remedial math project was successful, we are, in essence, saying that our children were like 20th percentile children before the intervention and like 30th percentile children after the intervention; therefore the intervention was effective in raising their performance level. Put a slightly different way, we assume that our children would still be like 20th percentile children at posttest time if they had not participated in the project but, because they did, their performance has improved. A little thought will quickly reveal that we are using the publisher's normative data in exactly the same way we would use data from a control group. It is this analogous usage that we must keep in mind when implementing a norm-referenced evaluation or we are likely to make possibly serious errors.

To illustrate the problem, the following short quiz has been devised:

1. A group of 100 sixth graders are randomly assigned to treatment and control groups. The treatment group is pretested on October 15th. When should the control group be pretested?

   a. September 15th
   b. October 15th
   c. November 15th
   d. some other time

2. The treatment group is posttested on May 1st. When should the control group be posttested?

   a. April 1st
   b. May 1st
   c. June 1st
   d. some other time

3. If the control group had been pretested September 15th and posttested June 1st, the experiment would

   a. overestimate the treatment effect
   b. underestimate the treatment effect
   c. correctly estimate the treatment effect
   d. be inconclusive

4. If the control group had been pretested November 15th and posttested April 1st, the experiment would

   a. underestimate the treatment effect
   b. overestimate the treatment effect
   c. correctly estimate the treatment effect
   d. be inconclusive

5. If the control group had been pretested September 15th and posttested April 1st, the experiment would

   a. underestimate the treatment effect
   b. overestimate the treatment effect
   c. correctly estimate the treatment effect
   d. be inconclusive

The correct answers to the above questions are b, b, b, b, and d. The only time when the experiment can correctly estimate the treatment effect is when the treatment and control groups are tested at the same time. Of course, if the

control group is pretested a month earlier and posttested a month later than the treatment group, the comparison will be unfair and it is almost certainly safe to assume that the treatment effect will be underestimated. Conversely, when the control group is pre- and posttested over a shorter time interval than the treatment group, it is safe to assume that the treatment effect will be overestimated. When the pre-to-posttest interval is equal for the two groups but the beginning and ending points do not match, it is impossible to tell whether the treatment effect will be over, under, or correctly estimated.

The need to test both groups at the same time applies just as strongly to norm-referenced evaluations as to control group studies. When implementing this type of evaluation, it is essential to determine when the test was standardized and to test children in the evaluation sample at, and only at, that time. The fact that test publishers provide norms tables for fall, winter, and spring and claim that those norms are each "good" for three calendar months is quite misleading. Such norms will yield only approximate values except at times that correspond to empirical norming dates--and approximate values are not good enough when attempting to estimate treatment effects.

The comparable testing time requirement applies to all evaluation designs that make use of empirical measurements. If test scores from the year preceding the intervention are used as the without-intervention expectation, it is just as important to observe this requirement as when control group or norm-referenced models are implemented.

For example, the Vocational Education Amendments of 1976 require yearly reports on the post-program status of graduates: enrollment in further education, employment in training-related occupations, and employer satisfaction. College enrollment is usually highest in September; youth unemployment is usually highest in summer. If the innovative vocational education is to avoid looking better or worse than it actually is, comparable times for data collection from one year to the next are essential. As another example, most students decide by the fall of their senior year whether they will go on to post-secondary education; most underclass students decide in spring what courses they will take during the coming year. Changes in guidance outcomes such as measured certainty of choices in career planning may be susceptible

to fluctuations from fall to spring and may also interact with educational status.

It is occasionally possible to conduct a credible evaluation without an empirically based estimate of what conditions would have been without the intervention. Where a totally new topic is introduced into a curriculum, for example, it is not necessary to demonstrate empirically that a control group, not exposed to the topic, learned less of it. If the treatment group learned anything as a result of the new instruction, it could clearly be expected to outperform the control group and no proof would be required. On the other hand, it should be demonstrated that the new learning was in addition to, not in place of, what was learned previously.

Care must be exercised not to overgeneralize regarding the applicability of logic as a substitute for empirical data. There is an unfortunate tendency in many circles to do just that when dealing with objectives-referenced instruction and testing. Simply demonstrating that 90 percent of the students "mastered" 90 percent of some set of objectives is not adequate proof of effective instruction since it is at least possible that 90 percent of a control group would also have mastered 90 percent of the same objectives without the special intervention. Unless such possibilities can be convincingly dispelled one way or another, the evaluation results must be regarded as inadequate. Again, there must be a credible estimate of what conditions would be like without the intervention.

Conclusions

The major thrust of this chapter has been to describe ways of making a convincing case that educational interventions have produced change. The first point made was that care needs to be exercised in identifying and describing important outcomes that are directly and logically related to the objectives of the intervention. While failure of a JDRP submission to distinguish between processes and outcomes obviously does not reflect any lack of success of the intervention, it may significantly complicate the JDRP review. It could, in extreme cases, so obscure the relevant information that a truly successful project might be rejected.

The remainder of the chapter dealt with compiling and presenting evidence that changes had, indeed, occurred. It was pointed out that convincing evidence must always include explicit or implicit comparisons between conditions that exist after the intervention and an estimate of what those conditions would have been without the intervention. Unacceptable and acceptable evaluation designs were discussed along with the conditions under which each of the latter would yield credible results.

It should have been apparent from the discussion that there is no completely satisfactory way to conduct an evaluation and that no evidence can ever be completely convincing. For this reason, an effective strategy is to assemble as many different kinds of relevant evidence as possible. The combination of several pieces of moderately convincing evidence may be more than sufficient. The situation is similar to criminal justice where the testimony of a single eye-witness to a crime would probably not be considered sufficient evidence for a jury to convict the alleged perpetrator. However, this testimony, when combined with fingerprints found at the scene of the crime and possession of the stolen goods, would be enough to sway even the most skeptical juror.

A Convincing Example

The first convincing example is drawn from one of the earliest Dissemination Review Panel submissions, Project Catch-Up of the Newport-Mesa Unified School District in California. This project was approved in early 1974, before the DRP became the JDRP.

The submission was prepared by an independent contractor not involved in the design or execution of the original evaluations. It thus involved reanalysis of results judged to be inconclusive as originally presented. The following caveat was included: "Unfortunately, the test instruments used precluded direct application of the most powerful analytical models and forced the adoption of unduly conservative alternatives. While these conservative analyses conclusively supported the success of Project Catch-Up, the findings would almost certainly have been more impressive had it been possible to employ more direct evaluation models."

The evaluators used a quasi-experimental design, in which student performance was compared with national norms. The Cooperative Primary Test (which has empirical norms at 1.7, 2.1, 2.7, 3.1, and 3.7) was used in first, second, and third grades in 1971-1972 (sample sizes were not specified). Pre-to-posttest gains of 13.5, 20.2, and 10.3 percentile ranks were reported for respective grade levels in reading and 16.9, 11.7, and 23.4 in math (although the first-grade results should be discounted as empirical norms are not available for beginning of first grade).

In 1972-1973, the Reading scale of the Metropolitan Achievement Test was used in grades one, two, and three. Unfortunately, gains were reported in terms of grade equivalents, and individual student raw score data could be retrieved only for the third graders (N = 38). Both pre- and posttesting were accomplished acceptably close to empirical normative data points and a very large gain of 31 percentile points (1.16 standard deviations) was found. Cooperative Primary reading scores were located for 44 second graders pretested in spring of first grade and posttested in spring of second grade. These children were observed to move from the 14th to the 30th percentile--a gain of .55 standard deviation.

In grades four, five, and six, the Comprehensive Test of Basic Skills (Form Q) was used on a fall-to-spring testing schedule in both 1971-1972 and 1972-1973. There were two fundamental defects in the evaluation: grade-equivalent scores were used, and testing was accomplished at times far removed from empirical normative data points. Nevertheless, some interpretation was possible.

In reading in 1971-1972, the 35 beginning fourth graders (grade 4.1) scored 80.5 standard score points (1.18 standard deviations) below the mean of children at grade 3.6 at pretest time. At posttest time (grade 4.8), they scored 20.5 standard score points (.28 standard deviation) below the mean of children at grade 4.6. They thus showed a gain .9 standard deviation larger in seven months than was made by the national norm in twelve months. While the gains were substantially smaller for the 37 fifth graders (.22 standard deviation) and 17 sixth-grade children (.24 standard deviation), they still exceeded yearly expectations over the

seven-month period. Gains in math were slightly lower than those in reading at each corresponding grade level.

In 1972-1973, fall-to-spring gains in reading exceeded yearly expectations by .34 standard deviation in fourth grade (N = 49), by .59 standard deviation in fifth grade (N = 34), and by .17 standard deviation in sixth grade (N unspecified). The corresponding gains in math were .45 standard deviation in fourth grade and .44 standard deviation in fifth grade. Data for the sixth graders were uninterpretable as they had been given an abbreviated version of the test.

Through sheer weight of evidence, this example builds up a convincing case for project effectiveness. It should be noted that the evaluators avoided the pitfall of simply testing the statistical significance of growth from pre- to posttest. The educational significance of the gains made by project participants was established through comparison with gains observed in the national norm group, expressed in terms of percentiles and standard deviations.

Another Convincing Example

A second, totally different kind of example comes from the Diversified Educational Experience Program (Project DEEP) of Wichita, Kansas. The goals of this project were non-cognitive, and the evaluation did not use standardized tests or any other measures of academic achievement. The project was approved by a JDRP vote of seven-to-one in 1976.

The overall purpose of Project DEEP was to improve the attitudes and behavior of secondary students in schools where dropouts, absenteeism, and poor attitudes were common problems. The success of this Title III program was measured in terms of reduced absenteeism, lowered dropout rate, and miscellaneous other indices.

Project evaluators used a control-group model. A process whereby students were randomly assigned to classes, and classes were randomly selected for participation in the program, effectively eliminated any systematic differences between the 711 DEEP students and the 6975 non-DEEP students in the schools. Comparisons between the two groups of students at the end of the 1972-1973 school year showed 30 percent fewer absences and a 37 percent lower dropout rate for

the DEEP students. Chi-square tests showed these differences to be statistically significant at the .01 and .001 levels, respectively.

Further evidence of success included the Affection, Respect, and Knowledge scales of the Risk Taking Attitude Values Inventory, all of which showed gains from pretest to posttest that were statistically significant beyond the .05 level.

In addition to the original control-group evaluation, a follow-up study was performed that provided two additional items of evidence supporting the success of Project DEEP. Ninety-eight percent of Project DEEP juniors subsequently graduated, while the graduation rate for non-DEEP juniors ranged from 91 percent to 94 percent in the Wichita Public Schools.

Examples of Unconvincing Evidence

A Title I reading program had both pre- and posttest scores on Form E of the SRA Primary I Reading Test for 3,623 second-grade students. The mean number of items answered correctly on the mid-October pretest was 15.0. At posttest time in mid-April, the mean performance was 20.0 items answered correctly. This difference was tested using a *t*-test for correlated data and was found to be statistically significant at the .001 level ($t = 3.38$, $df = 3620$).

The grade-equivalent scores corresponding to the mean pre- and posttest raw scores were 1.1 and 1.9, respectively, showing that the students had made an eight-month gain in the six-month interval between testings. Therefore, the claim was made that the participating students had made gains of one-and-a-third months per month and that the program was successful.

The statistically significant raw score gain had to be judged irrelevant to the evaluation as second graders are expected to score higher at the end of the year than at the beginning, whether or not they participated in a special project. The grade-equivalent gains had to be dismissed because of the inconsistencies known to exist in this metric.

A check of the SRA norms table (revised in fall 1974) revealed that both the pretest and postest mean raw scores

corresponded to performance at the 21st percentile. This indication suggested that the children had not improved their status with respect to the national norms and, consequently, had not benefitted from the project.

Although percentile gains are valid indices of student gains when tests are administered at times corresponding to empirical data points, the SRA test currently has normative data for spring (mid-April) only. The fall (pretest) percentile was thus based on data interpolation and must be considered too unreliable for a project-effectiveness measure. The only conclusion that could be reached, therefore, was that the available data supported neither the success nor the failure of the project.

In another project, a treatment was designed to enhance student self concept by providing opportunities for success experiences in academically related tasks. It was hoped that improved attitudes toward school would result and that the attitude improvement, in turn, would lead to improved academic performance. The project ran for three years, involving approximately 60 second, 60 third, and 60 fourth graders each year. Children who began in the first year as second graders were in the project for three years; those who began as third graders were in for two years, and the original fourth graders were in for one year. A new second-grade group was added each year as the original children moved up.

Teacher ratings of the project were quite high each year, ranging from a low of 7.3 (on a ten-point scale) the first year to a high of 7.6 the third year. The support of the administrative staff was also high and the school board voted to continue the program with local funds when the Federal funding expired. Unfortunately, there was no baseline against which to compare these data. Support could be high for an ineffective project.

The Coopersmith Self-Esteem Inventory was administered both fall and spring each year, and each year the spring scores were higher than the fall scores at each grade level. During the first project year, all three grade levels showed gains significant at the five-percent level; only second-grade gains were significant in the second and third years.

The evaluator noted that fall scores of children continuing in the project each year returned to approximately the same level as their fall scores before their first year of participation and attributed this over-the-summer loss to the removal of the reinforcing treatment. When questioned regarding the possibility that testing conditions might have differed from fall to spring, he responded that he had instructed the testers to follow exactly the same procedures at each test administration. He acknowledged, however, that the tests were given by the project teachers, that they might be ego-involved in the success of the project, and that he had not observed the test administration. Since neither this nor other plausible alternative hypotheses (e.g., everything looks better just before a vacation) could be rejected, the self-esteem data were regarded as unconvincing.

No gains were shown on reading or math tests that exceeded those observed in years prior to project implementation. School staff were sure that attendance had improved but, unfortunately, attendance records for years prior to project implementation had been lost when the administrative offices of the school were renovated.

Parent attitudes toward the program were positive, but only mildly so, with the median response falling into the Slightly Improved category of the Attitude Toward School Inventory. Seventeen percent of the parents reported that they were unaware of the project's existence.

IV. CRITERION 2: WAS THE EFFECT CONSISTENT ENOUGH AND OBSERVED OFTEN ENOUGH TO BE STATISTICALLY SIGNIFICANT?

We believe that the sun will rise in the east tomorrow because it always has in the past. We make Jack Nicklaus the favorite to win the U.S. Open because he has won more golf tournaments than any other player on the tour. We will draw a card if we hold two pair but will not risk drawing to an inside straight. And we will put a coin in the slot machine even though we know the odds favor the house. These are all examples of how we behave when confronted with varying degrees of uncertainty.

If we conducted an educational experiment and found that twelve of the 30 participating children did a little better on the posttest than would have been expected without the experiment while nine did about as expected and nine did a little worse, we would not be very confident that the experiment was a success. On the other hand, if we repeated the experiment ten times and got the following results--

TABLE 1. RESULTS OF EXPERIMENT REPEATED TEN TIMES

Experiment Repetition	Direction of Change on Posttest		
	+	0	-
1	12	9	9
2	13	8	9
3	14	8	8
4	15	8	7
5	11	9	10
6	13	5	12
7	9	13	8
8	11	10	9
9	9	14	7
10	13	7	10

we might be quite confident that the experiment had a positive, although small, result since all ten of the experiments showed positive impacts. (It should be noted that the average of the ten experiments is exactly the same--12, 9, and 9--as the result of the first experiment, which we judged ambiguous.)

Had the results of our original educational experiment showed that 28 of the children scored higher than expected, one about the same as expected, and one lower than expected, we would have been far less reluctant to claim that the experiment was a success. The 93 percent rate of improvement is convincing despite the small number of students involved. Statistically, in fact, this latter single experiment is far more significant than the sum total of the ten experiments in the earlier example (which, however, is also significant at the .05 level).

If the preceding examples have served their purpose, they will have illustrated several points: (a) small gains observed in small samples are neither statistically significant nor intuitively convincing, (b) small gains become statistically significant and intuitively convincing if observed often and consistently enough, and (c) large gains are both statistically significant and intuitively convincing with fewer observations. What should be noted from these observations is that there is a high degree of correspondence between statistical significance and what one would infer from simply examining the data. The one exception might be the case where a very small gain was statistically significant because it was observed in a very large number of cases--and, of course, such a small gain would lack educational significance.

Statistical tests assume that non-treatment-related variations in the data are random. In real-world educational evaluations, however, this assumption is probably never quite met. It is possible that biases may be present (too much time allowed on the posttest, not enough on the pretest; guessing encouraged on the posttest, discouraged on the pretest; a disruption occurred during the control group's testing session, etc.). Such biases would spuriously inflate the apparent statistical significance of the results, but they would also negate their credibility.

For this reason, few evaluators would unquestioningly accept the results of any single, small-scale study as adequate evidence of the success of any intervention regardless of the level of statistical significance. There should always be at least one replication or parallel study. If, for example, comparable results are observed in two or more classrooms, or in two or more successive years, or both,

results become much more credible. This type of consistency of findings not only helps to establish statistical significance and intuitive credibility, it is also directly relevant to the transportability criterion and, less directly, to the other criteria discussed in subsequent chapters.

In presenting evidence regarding the frequency and consistency of observed effects, it is not necessary to employ sophisticated statistical techniques. While analyses based on covariance or regression techniques are perfectly acceptable and may provide insights into educational phenomena beyond those provided by simpler analyses, they are no more convincing than more straightforward treatments of the data. In fact, when such techniques are necessary simply to demonstrate that an intervention had some impact, it is likely that the impact was small, that there were shortcomings in the evaluation design, or that some combination of these two was precluding a more direct analysis.

While complex statistical treatments of data are neither required nor encouraged, few evaluations will yield data that cannot be subjected to some form of statistical test. (One such example, however, is presented in the next chapter.) The second "convincing example" in the preceding chapter, for instance, tested simple frequency counts of dropouts and non-dropouts for treatment and comparison groups. The data are presented in Table 2.

TABLE 2. NUMBER OF DROPOUTS AND NON-DROPOUTS IN EXPERIMENTAL AND CONTROL GROUPS

	Dropouts	Non-dropouts
Experimental (DEEP)	57	654
Control (non-DEEP)	886	6089

A simple chi-square test was performed from which the dropout rate was found to be significantly smaller for the project participants at the .001 level of confidence, and a very convincing case was made. While the project might not have been approved had that been the only evidence presented, none of the analyses included in the submission was any more complex.

Similar analyses could be done comparing children exposed to some intervention with those not exposed in terms

of number of objectives mastered and number not mastered on a criterion-referenced test. If a concurrent control group cannot be found, perhaps historical or normative data could be used. In any case, the single most important thing to remember is to present data on as many instances as possible. If the same intervention was tried in several classrooms in several grade levels in several schools, all of the evaluation data should be presented.

One final point needs to be made. The results of all statistical tests are expressed in terms of the probability that the observed results could have occurred by chance. When $p < .05$, that means that the obtained results would be expected to occur by chance alone in fewer than five out of one hundred similar situations. It follows logically from this relationship that when 20 comparisons are made between treatment and comparison group students, one would reach the .05 level of statistical significance by chance alone. Thus, evaluations that report one out of twenty, two out of forty, or four out of eighty comparisons as reaching the .05 level of statistical significance provide no evidence that the intervention was successful.

A Convincing Example

The 1974 submission of Project R-3, a junior high school program in San Jose, California, contained both positive and negative results. It encompassed, however, evaluation results for a five-year period from the 1967-1968 school year through the 1972-1973 school year. For most years, data were presented in both reading and math and often for two or three grade levels.

Some of the evaluations used control groups while others were norm-referenced. Some employed grade-equivalent scores and were uninterpretable. In other instances the analyses were conducted using both grade equivalents and more appropriate metrics. It was often the case that, for given grade levels in particular years, gains did not exceed expectations in both reading and math. The overall pattern was clear, however. In most years, at most grade levels, statistically significant gains were made either in reading or math or both. The Panel approved the project unanimously although, had the submission been based on data from only one or two

years, the pattern of success would have been less apparent and the project might not have been approved.

A Second Convincing Example

Project MECCA compared two types of programs for the prevention and remediation of specific learning disabilities of kindergarten children. MECCA involved early detection coupled with collaboration between the learning disability specialist and the regular classroom teacher. The traditional program provided screening, referral, and "pull out" for special tutoring. The evaluation design was replicated in two school systems, and within each school system in two MECCA and two traditional schools. The Jansky Screening Index was administered at the beginning of the study to identify children within the high-risk range of learning disabilaties. The Jansky and the Metropolitan Readiness Test were administered as posttests. Analyses of the data showed:

- no pre-treatment differences on the Jansky
- better post-treatment scores for MECCA students ($p < .01$)
- higher Metropolitan scores for the MECCA children ($p < .001$)

Additional supporting evidence included fewer MECCA than traditional-program children referred for special education in the first grade when screened by school system psychologists ($p < .05$), and, a year later, a high percentage (74 percent) of MECCA children scoring above the high risk level of the Jansky ($p < .05$). Again, a convincing case was made by accumulating statistically reliable effects in several sites for both immediate and longer term outcomes, rather than by a single, overwhelmingly large or reliable difference.

Unconvincing Example

One Title I project attempted to improve the acquisition of reading skills through the matching of instructional treatments to the learning styles of the participating students. Seventy-two children were assigned to either a "linguistic" ($N = 38$) or a basal-reader treatment ($N = 34$) based on their responses to a specially developed learning style

test. Performance was measured at the end of the school year using the Stanford Achievement Test Word Meaning and Paragraph Meaning subtests.

Both groups were observed to make approximately month-for-month grade-equivalent gains and, since they were well below average readers, this finding was offered as evidence of success.

An additional analysis was done by sex and revealed that the boys significantly outperformed the girls in the basal-reader program (making 1.4 months gain per month as opposed to .8 month per month for the girls) while the girls significantly outperformed the boys in the linguistic treatment (1.5 months per month as opposed to .7 month per month). It was argued, following the presentation, that the effectiveness of the program could be further enhanced by assigning males to the basal-reader group and girls to the linguistic group.

There are several problems with this example. First, the original conclusion based on grade-equivalent scores is not convincing. Second, the argument regarding sex differences is questionable at best, since the assignment of both boys and girls to treatment groups was based on their scores on a learning style test whereas it should have been random. With nonrandom assignment, the children of each sex who received one treatment were systematically different from the children of the same sex who received the other. Thus, the experiment generated no evidence regarding how either the particular boys or the girls who were selected to receive one treatment would have done if they had been given the other.

The third problem with the example is that it was a one-shot study involving only a small number of students.

Any one of the three defects in this example would have been sufficient for JDRP rejection of the project.

V. CRITERION 3: WAS THE EFFECT EDUCATIONALLY SIGNIFICANT?

The issue of educational significance, unlike that of statistical significance, is not a technical question. It has been termed a matter of common sense, but the sense is often not common--that is, different judges cannot agree. In any case, educational significance is a matter of judgment.

The job of the JDRP would be an easy one if the typical submission described interventions that produced large effects, that required minimum expenditures, and that entailed minimum disruption of ongoing educational programs. Unfortunately, such submissions are very rare. More typically, effects are small, costs are large, and the issue of educational significance is not at all clear. The following discussion focuses on these marginal cases.

In judging the educational significance of an intervention's impact, two factors must be considered: the size of the effect and the importance of the area in which it happened. There ought to be a reasonable balance between the two factors. The chance that a small gain would be considered educationally significant is higher in a broad or educationally important area than in a narrow or less important area.

There is a parallel between the breadth of focus of educational interventions and that of the measures used to assess their impact. Standardized achievement test scores are the most widely known and accepted measures for use in educational evaluations. They have, however, been justifiably criticized on several counts. From an evaluator's viewpoint, the most relevant criticism is that they do not measure what is being taught. On the other hand, most people concede that such instruments do measure the ability to read, do arithmetic, etc., issues of racial or ethnic bias notwithstanding.

Standardized achievement tests are designed to be curriculum-free. Only to the extent that they achieve that objective can they be employed to assess the skills of children exposed to different instructional programs. Being curriculum-free, on the other hand, cannot mean curriculum-irrelevant. A better term might be curriculum-fair--a label

that would imply that each existing curriculum would be represented by an equal number of items in the test.

Curriculum fairness represents some sort of ideal that can never be attained. In striving for it, however, test publishers end up producing instruments that contain a relatively small proportion of items that are directly taught in any particular curriculum. For this reason, they are less sensitive to the treatment effects of any specific program of instruction than they would be if composed solely of curriculum-relevant items.

It is the generality of standardized achievement tests that lends credibility to the educational significance of any gains that can be shown on them. This same characteristic, however, makes such tests relatively insensitive to the kind of gains produced by many types of educational interventions. They will almost certainly fail to show whatever gains may result from short-term, highly focused interventions.

Objectives- or criterion-referenced tests will be far more sensitive. They, however, may suffer from credibility problems when questions are raised regarding the educational significance of gains. If, for example, the results of an evaluation reveal that treatment-group children averaged five more items correct on a criterion-referenced phonics test than children in a control group, skeptics are likely to say, "Yes, but did it help them learn to read?"

Despite this source of difficulty with criterion-referenced tests, it is important to reiterate that the sensitivity of any test to measure change will be directly related to the relevance of the test items to the content of the instruction. It should also be pointed out that there <u>are</u> content differences among standardized achievement tests. Examination of the items that make up each test will enable the identification of the test that is most relevant. This test, in turn, will be the test most sensitive to whatever gains are produced by the intervention.

It is, of course, not necessary to employ achievement tests of any type in order to establish the effectiveness of an educational intervention. The most convincing evidence of success in a dropout-prevention program, for example,

would simply be statistics showing a decrease in the number of students dropping out. Similarly, vocational education programs might be evaluated in terms of job placements, starting salaries, rates of advancement on the job, etc., although achievement tests could also be used.

Assuming that measures have been selected that are appropriate for and sensitive to the outcomes of any particular educational intervention, the question of educational significance must still be resolved in terms of the size and importance of observed effects. Consider two examples: (a) a reading program that, during the course of an entire school year, succeeded in raising the status of the participating children twelve percentile points in Total Reading with respect to the national norms; and (b) a phonics program that spent an entire year exclusively on diphthongs and produced a similar gain with respect to the national norms on a test of diphthongs.

Clearly, the first of these two outcomes would be considered more educationally significant than the latter. Even the latter, however, might be considered educationally significant if it had been achieved at a very low cost and without interfering with other learning. Cost, in fact, can never be totally divorced from the consideration of educational significance. Assuming that resources are limited, it is clear that more children can be served by low-cost interventions than by high-cost interventions. If the gains are comparable, the fact that the low-cost intervention serves more children makes it more educationally significant as well.

While the concept of cost effectiveness is relevant, it must not be taken as the sole criterion for educational significance. Interventions that work in some settings for pupils with particular needs will not necessarily work in other settings or for pupils with somewhat different needs. It would probably be quite inappropriate, for example, to make cost-effectiveness comparisons among projects serving (a) urban ghetto children, (b) bilingual migrant children, (c) visually handicapped children, (d) dyslectic children, and (e) institutionalized delinquent children.

To summarize, an educationally significant effect is one of nontrivial magnitude, in a content area generally

accepted as important, which can be achieved at a reasonable cost. The fact that these three dimensions are all interdependent makes the establishment of strict cutoff values impossible. The further complication that interventions serving different types of children in different settings can be expected to have different sized impacts at different costs makes the decision-making process an extremely complex judgmental matter.

Theory, past experience, expert judgments, and statistical rules of thumb have been invoked in weighing the size of educational effects.

Past experience has been applied in evaluations of career education and a compensatory reading program. In both instances, the literature was combed to summarize the frequency, direction, and size of differences found between experimental and control groups in previous studies. It was then decided that projects that showed effects comparable to those of the upper 10-20 percent of previous, positive studies would be regarded as educationally meaningful.

Expert judgments and theory. The magnitude of effects of a program intended to increase the self reliance of severely retarded institutionalized youth might be judged by experts in institutional care. The progress made in dressing, toiletting, and feeding might seem small to those accustomed to even slow-learning children. Those familiar with the usual status of institutionalized retarded individuals might regard the improvements as highly worthwhile. In an actual instance, inner-city high school principals were asked to establish educationally meaningful school retention rates. New techniques for a priori determination of the social utility of outcomes should lend more power to expert or stakeholder judgment as a criterion of the educational meaningfulness of results.

Statistical rules of thumb. One widely applied rule is that the effect must equal or exceed some proportion of a standard deviation--usually one-third, but at times as small as one-fourth--to be considered educationally significant. Another statistical criterion is a rate of growth that will produce a posttest percentile standing that exceeds the pretest percentile standing by one standard error of measurement.

In another example, the percents of experimental and control children still in special education classes ten years after an early childhood intervention were compared to cumulative special education costs. "Payback" was reached after only a few years. Thus, even with the fairly small absolute differences, the program effect was judged educationally significant.

The examples presented below may be somewhat controversial. Unfortunately, while noncontroversial examples would be easy to find, they would not be particularly helpful. The examples were deliberately chosen to illustrate the complexity of the judgments involved.

A Convincing Example

A geographically extensive rural school district leased a van and outfitted it as an optometric test center. An optometrist and an assistant traveled from school to school in the district checking the eyesight of participants in Title I projects. Nearly 5,000 children in kindergarten through grade twelve were examined during the 1974-1975 school year. One hundred and twenty-six children were identified who were judged to have serious visual impairments that were previously undiagnosed and uncorrected. An additional 211 children were identified whose vision was judged significantly impaired.

The impairments of all but four of the children in the "severe" and "significant" categories could be corrected to within the "normal range" with proper refraction. Eyeglasses were prescribed, fitted, and delivered to the needy 333 students. The four students with more serious problems were referred to the appropriate county health service agencies for further diagnosis and treatment.

The initial impression was that the educational significance of this treatment was essentially unassailable, even though unsupported by statistical tests. The visual impairments were clearly of sufficient magnitude to interfere with classroom learning, the treatment was delivered to the neediest children, and presumably, the treatment was effective in correcting the problem. When costs were considered, however, a question arose as to whether the same effect could not have been more economically achieved in some other manner. Busing children to county clinics where similar services

could be performed was specifically considered as an alternate approach. Analysis revealed, however, that the cost of this alternative was about equal to that of the mobile van, and it was judged that the mobile van was far less interruptive of ongoing educational processes.

A Second Convincing Example

The Electric Company television program was approved by the JDRP based on an independent, contractor-conducted evaluation in Fresno, California. The design involved matched experimental and control classrooms and about 3,000 students in grades one through four during the 1971-1972 school year. Tests given both pre and post included:

1. The Electric Battery--a 123-item test designed to test students' performance on the program's curriculum objectives. It yields subtest scores for "matching words," "blending," "chunking," "scanning," and "reading for meaning."

2. The Word Knowledge subtest of the Metropolitan Achievement Test.

The following table presents the *t*-values obtained by comparing treatment- and control-group posttest performance measures.

TABLE 3. RESULTS OF *t*-TESTS COMPARING TREATMENT AND CONTROL GROUPS BY GRADE

Test	Grade 1	Grade 2	Grade 3	Grade 4
The Electric Battery:				
Matching Words	-1.32	1.87*	-.27	.08
Blending Total	1.75	1.99*	.86	1.38
Chunking Total	3.11*	4.26*	2.14*	2.59*
Scanning Total	2.56*	4.30*	2.98*	2.45*
Reading for Meaning	1.40	2.46*	1.54	1.04
Grand Total	2.20*	3.56*	2.37*	2.03*
MAT--Word Knowledge	.73	.38	.82	2.15*

*One-tailed statistical significance at the .05 level.

The results were somewhat ambiguous with only two subtests and the total score for the Electric Battery showing statistically significant differences across all four grade levels. The size of the t-values, furthermore, suggested that with 3,000 cases, the between-group differences must have been rather small. On the other hand, the cost per pupil was estimated to be only $1.20. Even a modest gain achieved at so little cost is impressive!

A Third Convincing Example

An alternative school for inner city youth 15 to 21 years of age, half of whom had already dropped out of school, based its curriculum on infusing career education into an intensified academic curriculum. As compared with true controls selected at random from eligible applicants, the Career Intern Program claimed greater school retention, improved basic skills, and improved career planning skills. As shown in Table 4, the gains made by the interns (N = 88) were statistically larger than those made by the controls (N = 55) except for a measure of self-esteem. They were also judged educationally significant. The per-pupil costs were higher ($2,384) than the general and academic tracks ($1,615), the same as the vocational education program ($2,314), and lower than tuition in a private school ($2,784). Because of the magnitude of the effects, coupled with their statistical reliability, the project was judged exemplary even though per-pupil costs were quite high.

Unconvincing Examples

A project in a large metropolitan area was designed to teach the principles and practices of ecology. The course was taught in kindergarten through grades four, six, and ten, replacing the regular science component of the curriculum in those grades.

Gains were measured using locally developed achievement (criterion-referenced) and attitude scales. Posttest minus pretest raw scores were computed for individual students and pretest-to-posttest gain comparisons were made between treatment and control groups. The achievement gains for both groups were rather small, ranging from -.6 for the tenth-grade controls to 6.0 for the third-grade treatment group and averaging 3.3 points for the treatment groups and .9 for

TABLE 4. STATISTICS COMPARING PRETEST AND POSTTEST FOR TREATMENT (INTERNS) AND CONTROL GROUPS

	Pretest X	Pretest SD	Posttest X	Posttest SD	p
Stanford Achievement Test					
Reading					
Interns	28.5	13.5	30.6	14.8	<.01
Controls	28.4	11.9	24.5	17.6	
Math					
Interns	14.4	8.1	16.0	8.2	<.01
Controls	13.6	9.1	12.3	9.0	
Super Career Development Inventory					
Resources					
Interns	256.8	54.0	298.9	58.1	<.01
Controls	252.3	42.6	261.3	53.9	
Information					
Interns	13.1	4.6	16.4	5.3	<.01
Controls	12.7	4.5	13.8	4.5	
Planning					
Interns	96.9	22.1	122.9	21.5	<.001
Controls	96.8	16.6	108.2	20.8	
Coopersmith					
Academic Self Esteem					
Interns	4.1	1.7	5.4	2.0	<.05
Controls	3.7	1.6	5.1	1.8	

Retention	Graduated or in school	Drop out
Interns	69%	31%
Controls	23%	77%

the controls. In first grade, the treatment-group children averaged only 1.4 points higher on the 21-item test at posttest time than they did at pretest time.

Although differences between treatment and control groups were statistically significant (except at kindergarten, where $p \leq .06$), they were small, ranging from 1.0 item in first grade to 4.7 items in third. More importantly, the test that was used to make these comparisons had been developed locally by successively discarding items (from an original pool) that failed to discriminate between project and control students. JDRP members found the gains not only small but uninterpretable.

In another example, a computer-assisted instruction project in one of our major cities was responsible for raising the reading achievement test scores of 108 Title I fourth graders from the 15th to the 21st percentile of the national norms between an October pretest and a May posttest using the Sequential Tests of Educational Progress. Although the evaluation sample was small and the study had not been replicated, the results were credible because of very carefully controlled experimental conditions. Per-pupil expenditures, however, exceeded $1,900. It was clear that the gains could not have been replicated at lower cost since the major item of expense was the very sophisticated terminals used by the project and utilization of the terminals was nearly 100% during regular school hours.

VI. CRITERION 4: CAN THE INTERVENTION BE IMPLEMENTED IN ANOTHER LOCATION WITH A REASONABLE EXPECTATION OF COMPARABLE IMPACT?

An intervention that worked once in a given location, with a certain group of staff members, and with a particular group of participants, may be effective in the event that the original circumstances are changed--or it may not. This point is easy to overlook in assessing the effectiveness of products or practices, but it is as important in education as in other fields. New drugs that appear promising in the laboratory are still given field tests with samples from the ordinary population before being released for widespread use because it is important to have some indication that they will work well outside of the laboratory. In the same manner, educational interventions should offer some degree of certainty that their effects will be the same when used with a new set of participants in a different setting. Statistical significance may reassure us that project results were no fluke, but that still does not guarantee that the intervention will be effective when replicated in other settings. In order to determine the likelihood that the same products or practices, when used elsewhere, will produce results similar to those obtained at the original site, the panel considers the following questions:

Is the Project Setting Unique?

An intervention that functions in an environment seldom found elsewhere may be deriving its effectiveness solely from that environment; in another location, the same products or practices may not work. For instance, an educational project could depend on surroundings or facilities that are not generally available. Suppose an auto mechanics course was developed at a high school that happened to be next door to a wrecking yard, where each student was allowed to work on a car of his own. One would not necessarily expect the program to be equally successful, even though all its instructional techniques were followed faithfully, in a high school that did not have equal access to old cars on which students could practice.

Is the Project Effect Solely Due to the Unique Characteristics of the Staff?

Most people can remember a teacher who inspired them to do well in a certain subject, even though the same material, taught by someone else, might seem difficult and tedious. This common experience is important to remember when judging the effectiveness of educational interventions. Is it the product or the instructional technique that produced impressive achievement gains, or is it Mr. Smedley, who taught the class? One way of guarding against the "Mr. Smedley effect" is to look for evidence from more than one class. Extremely small projects may have to provide data from more than one year before it can be established that the characteristics of the intervention, not of the teacher, are responsible for the impact. The importance of staff characteristics is not restricted to teachers; an extraordinarily dynamic principal or project director may produce the same kind of effect. Wherever it appears likely that one or more rare individuals exerted an influence that typical school personnel could not duplicate, the replicability of the intervention is in question. Unless submissions made to the JDRP describe interventions in sufficient detail, or unless their success has been achieved by more than one staff member, the possibility that the effect was due to charismatic leadership will remain in question.

What Evidence Is there to Suggest that the Intervention Would Work with Different Participants, in a Different Setting, and with a Different Staff?

Although the JDRP is concerned about the likelihood that an educational intervention will work equally well in another setting, few evaluations are designed to prove this. The primary focus is, as it should be, upon the effectiveness of the intervention as it was carried out. There are, however, various indications of generalizability that can be provided without going to the trouble and expense of conducting a replication study. One technique, as we have seen, is to involve more than one class and teacher wherever possible. Also, it is more convincing to use teachers who were selected randomly to implement a new approach than to have teachers who volunteered. The point is to provide evidence that the teachers who carried out the intervention were not unusual, so that one could expect teachers elsewhere to get similar

results if they used the same products or procedures. The same is true of other staff members--administrators, project directors, or specialists.

Similar considerations apply with regard to participants. The more there are, the better. Choosing them randomly produces a more convincing case for replicability than does using volunteers. If this is not possible, it is a good idea to collect any available evidence that will support claims that those who participated were not different from potential participants anywhere else, and that their performance was a typical result of the intervention, rather than a unique response to it by unusual participants.

One submission claimed a reading improvement program would be effective for all disadvantaged children, even though it had been tried out only for kindergarten through third grade, urban fringe, Spanish-speaking children. Universal claims were also made for curriculum modules in an ecology program dealing with water treatment and waste disposal even though the content was specific to urban northeastern conditions and the project only tested in these communities. Other submissions claim that all children (or teachers or parents or schools) are the intended beneficiaries, yet provide evidence only for far more restricted groups. Either moderating claims or presenting data commensurate with claims improves evidence of generalizability.

Lastly, is is important to examine whether the setting in which the intervention took place was unusual in some way that influenced the effects. Once again, if data from replications cannot be provided, a case for generalizability can at least be made by presenting evidence that similar settings can be found elsewhere.

What Components are Essential? Can These Be Disseminated?

Many educational innovations are compounds of commercially available curricula, project-developed manuals for in-service training, local strategies for parent involvement, individualization of instruction, and, almost always, special staff including the project director. The evidence about which a judgment of effectiveness is made reflects to a usually unidentifiable degree all of the above. Analysis of which components were really implemented (teacher training?

parent involvement? individualization of instruction?) can suggest, however, which components may be essential if the effects are to be obtained elsewhere.

As one example, evaluation of the the Career Intern Program used ethnography, case studies, interviews, and classroom observations to distinguish among components that were probably necessary and sufficient, those that were probably necessary but not sufficient, and those that were probably not necessary to obtain the observed outcomes. The evaluators also suggested which of the necessary conditions were non-replicable and which might be achieved through normal dissemination efforts. Although these assessments were not tested directly through replication, they afforded some sense of the likelihood of obtaining the effects elsewhere.

Examples of Convincing Evidence

Previous submissions to the JDRP and minutes of the panel meetings provide a number of examples of convincing evidence regarding the generalizability of an intervention. The most convincing evidence, obviously, would be data showing similar results from replication sites. Such evidence was presented by Project U-SAIL (Utah-System Approach to Individualized Learning), which submitted data from the original implementation site in Utah, together with additional data from three replications in Utah, one replication in Arizona, and one replication in Georgia.

Other examples illustrate, however, that expensive, large-scale studies are not necessary if there is solid, common-sense evidence that no unique factors other than the planned intervention caused the observed effects. For instance, the Math Laboratories for Disadvantaged Students in Honea Path, South Carolina, originally involved only two teachers. The obvious question was whether the observed success of the project was due to the presence of two outstanding teachers, rather than to the particular methods they employed. It was pointed out that, although the teachers were good, they were new teachers, both in their first year of teaching, and, consequently had not been selected for having outstanding teaching records. Such information can be convincing evidence that teachers in another location should be able to carry out the project with similar success.

In the case of a possibly unique project setting, replication studies are not the only way of providing evidence that the intervention can be expected to succeed in another location. For example, the "Learning to Read Through the Arts" project conducted by the Guggenheim Museum in New York City was approved when it was convincingly argued that its practices could be carried out in any city where there was an art museum or an art studio.

Examples of Unconvincing Evidence

One example of unconvincing evidence regarding difficult-to-replicate characteristics of the staff was a special tutoring project in which the tutors were student volunteers from a nearby college, all of whom were education majors. The submission claimed, however, that program effects could be obtained by any tutor and that no special qualifications were needed. This claim did not seem substantiated by the evidence since it was felt that these tutors might have substantially different skills from persons who might be available to fill their role in different settings.

Another project submitted survey results indicating that a majority of the teachers involved would be willing to teach in the program again. This raised some doubt as to how many teachers would not be willing to do so. Together with other doubtful factors, this left the panel unconvinced that the intervention had demonstrated success. It was felt that information on the acceptability of the project to the teachers, among other things, would need to be clearer before approval could be justified.

VII. CRITERION 5: HOW LIKELY IS IT THAT THE OBSERVED EFFECTS RESULTED FROM THE INTERVENTION?

It may be clear from the evidence presented to the panel that something happened--a given educational intervention definitely had an effect. But there remains the question of why something happened: Are other explanations of the effect equally plausible, or even more plausible? Since educational experiments can seldom be as rigidly controlled as experiments in the "hard" sciences, it is particularly important to ask this question.

Can Plausible Alternative Explanations Be Generated?

The JDRP considers evidence that improved performance is a result of certain educational interventions. Before one can feel confident that this is true, it is necessary to consider the possibility that other factors were actually responsible. The importance of proceeding with caution at this point can hardly be overemphasized.

The classical experimental design provides the most convincing approach to evaluation, but even it is not free of problems. The most important problem, of course, is that it is difficult to implement and often poses political, ethical, or even legal problems. In the case of ESEA Title I, for example, the regulations require that the neediest children be served. Thus, if it is possible to discriminate among children on the basis of need, it is clear that regulations disallow random assignment.

Even if the model could be properly implemented, potential problems would remain, and the findings of whatever data analyses were undertaken would require thoughtful interpretation. Consider, for example, the various possible impacts that may arise simply through knowledge of group membership. Children identified to participate in a Title I project may feel that they have been labeled as low achievers. They may feel resentful and thus be negatively motivated. On the other hand they may feel good about receiving special attention and consequently try harder (this phenomenon has been observed in many settings and is usually referred to as the Hawthorne effect).

Control group children may resent the fact that they are not receiving the same educational benefits as treatment group children, and this resentment may have a negative effect on their motivation. Or they may feel challenged to demonstrate that they can do just as well without the special treatment. (This phenomenon is sometimes called the John Henry effect.) All of these possibilities should be considered in interpreting evaluation results and, as far as possible, reasons should be presented why whatever gains were observed should be attributed to the treatment and not to other influences of the type just described. For further discussion of these issues, the interested reader is referred to Tallmadge and Horst (1976) and Campbell and Stanley (1963).

In addition to such side effects of experimental situations, there are often other reasons why real effects may be observed, but cannot necessarily be attributed to a special intervention. One reason might be the presence of other innovations that affected the participants at the same time. If a new math program is installed and, at the same time, every teacher in the school is given a full-time aide, it would be hard to prove that achievement gains in math were solely due to the new instructional program. Another confounding factor might be changes in the population that resulted in better performance after the intervention began. The growth of a new housing development within a district, or the redefining of district boundaries, could cause substantial changes in the school population. Achievement levels could appear to increase or decrease, even if a new instructional treatment that started at the same time had no true effect at all. Environmental factors affecting students can also lead to differences that may be wrongly attributed to an intervention. This problem affects control group designs in particular, and may occur whenever the treatment and control groups differ on any number of educationally relevant variables, for example, parental interest in education, exposure to books at home, presence of siblings, and so forth.

Still another explanation for observed effects might be normal growth of the participants. In cases where the initial status or the expected growth rate of participants has not been carefully established, the possibility that observed effects were simply the result of maturation is particularly troublesome. Successful compensatory education

projects, for example, may have effects that resemble normal growth for non-disadvantaged populations. In these instances it is important to provide evidence that the participants were really a lower-achieving group to begin with, and that the gains they made would not be expected simply from maturation, or from participation in a regular school program.

Can the Alternative Explanations Be Rejected?

In judging whether an educational intervention is effective, all of these possible alternative explanations--experimental side effects, other innovations, changes in the population, differences in factors affecting growth of two groups, and ordinary growth or maturation--must be considered. Only if all of them can be rejected should we feel confident that the intervention is in fact the cause of changes in performance.

For many of these alternative explanations, it is possible to make provisions in the design of the evaluation. For example, steps can be taken to prevent experimental side effects. Membership in the treatment or the control group need not be emphasized to either teachers or students. Finally, aside from factors characteristic of the intervention itself, conditions for the two groups should be as similar as possible. As usual, the bigger the groups involved in the evaluation, the better. Experimental side effects are less likely to show up across several schools or a whole district.

If a control or comparison group is used, every effort must be made to ensure that its members are as similar as possible to those in the treatment group. Systematic differences between groups in other factors such as urban or rural environment, race, socioeconomic status, or sex may be related to school performance. If these educationally relevant factors are not similar for both groups, it is difficult to determine whether the observed differences resulted from the intervention, or from differences in these other factors.

Some evaluation designs do not involve a comparison group; rather, the performance of the treatment group is compared with some norm or expectation. With this type of design, it is especially important to provide documentation of the initial status or the expected growth rate of the participants in the absence of the intervention. Such evidence

will make it possible to reject the possibility that normal growth or maturation might have been responsible for any gains.

Although a carefully designed evaluation may preclude some alternative explanations, such as experimental side effects or normal growth, it may not be possible to control for other circumstances in advance. Among such circumstances are events such as changes in the population or the introduction of other innovations at the same time as the intervention being evaluated. In some cases interference can be prevented; for example, close cooperation among all personnel in a district or school can ensure that several curriculum revisions are not undertaken at once if there is a possibility that one could affect the outcome of another. If simultaneous interventions cannot be avoided, then perhaps some form of control can be used. For example, in the example cited earlier of the math program and the teacher aides, an examination of student achievement in other areas affected by the aides, but not by any new curriculum, might help to isolate the reason for gains in math. If no such gains showed up in reading or social studies, one might assume that the instructional innovation, not the aides, had led to the improvement in math. It should also be possible to obtain documented evidence on such factors as stability of the population during the period of the intervention; this can rule out population changes as a cause of changes in performance.

Examples of Convincing Evidence

The problem of experimental side effects can be a very subtle one, difficult to isolate or measure. Nevertheless, "hard" data are not considered the only way to eliminate these explanations. As an illustration, the Diversified Educational Experience Program (DEEP) in Wichita, Kansas, had succeeded in reducing absenteeism and dropout rates in high school. Evaluators were asked "How do you know this isn't just a Hawthorne effect; maybe students just respond favorably to attention?" While no hard data were available on this point, one student was quoted as saying, "I'm not so sure I learned history, but I learned how to deal with the world, and I thank you." Such anecdotes, used sparingly, can lend credibility to the contention that what happened was real and not merely a psychological side effect.

Performance in other areas can serve as one indication of change or consistency in the student population. For example, a new reading program in the small town of Andover, Massachusetts, had apparently produced a significant improvement in the performance of the students on district-wide standardized reading tests. The question was whether the effects might have been the result of an influx of higher achieving students into the district. The evaluators stated that there had been no perceptible change in the composition of the population over the previous three years. To support this statement, they pointed out that the new program emphasized reading comprehension, and there had been large gains on the comprehension subtests. However, performance on word attack skills, emphasized in both the old and new programs, remained about the same before and after the intervention. It therefore appeared unlikely that the ability level of the students had changed.

Trend data offer another source of evidence that changes are attributable to the innovation. A "time out" room with academically intensive, individualized activities to which children with disciplinary problems reported was tried out as a way to reduce subsequent suspension/expulsion. The setting was a metropolitan area exhibiting urban decay. The year that this innovation was introduced, a dramatic reversal of a ten-year trend of increasing disciplinary problems, suspensions, and expulsions was observed. This reversal persisted for two successive years.

Examples of Unconvincing Evidence

Perhaps the most frequently encountered alternative explanation for the effects of an educational intervention is that they would have occurred anyway. In these instances, it cannot be established that the students involved in the intervention would not have gained just as much by another intervention, or even with no special treatment at all. One such case was a home-based educational program for early childhood. Data presented on a sample of five-year-olds compared the kindergarten performance of those who had participated in the early childhood program with the performance of those who had not. However, it appeared that those who had been in the program had come from more urban, upwardly mobile, and relatively advantaged homes. The logical conclusion was that the differences in performance reflected to

an unknown degree the continued influence of home factors independent of the program, as well as to some unknown degree an influence from the program.

Another alternative explanation that too often is not ruled out is testing condition bias. In one example, the teachers who established the objectives, designed the curricula, taught the children in the innovative programs, and wrote the tests also administered the tests pre and post. They also knew who were the participating children and who were not. Under these circumstances it could be expected that the project children would show greater fall-to-spring gains than the controls because of tester bias, teaching to the test, and greater familiarity of the children with the tester.

In contrast, another JDRP submission also used teachers as test administrators. In this evalution, however, for both fall and spring testing the teachers tested classes other than their own, helping to rule out tester bias as an explanation of the findings.

VIII. CRITERION 6: IS THE PRESENTED EVIDENCE BELIEVABLE AND INTERPRETABLE?

The preceding chapter dealt with cases where participants in an educational intervention made significant gains but where those gains may have arisen from causes unrelated to the intervention. A different type of problem occurs when it is not clear that benefits resulted for participants because flaws in the presented evidence cast doubt on this conclusion. There are a number of reasons why the data presented to demonstrate effectiveness may fail to do so, and the panel members must consider each of these when reviewing submissions.

Are There Any Apparent Inconsistencies in the Data Presented?

One of the most telling signs of a flawed evaluation is the presence of inconsistencies in the data. Sometimes they are obvious; in other instances they are more subtle and difficult to detect. A fairly obvious problem is lack of agreement between numbers reported in text and tables, or among tables. Another is apparent mistakes in the calculations. Other types of inconsistencies are more complex, and can occur in almost any phase of the evaluation. Information reported about how students were selected may not agree with pretest data, as for example when a compensatory education project claims to select students on the basis of need, but pretest scores show average-level performance. Or, the measures used may be clearly inappropriate for the purpose and content of the intervention, as when a mother-child program designed to increase the reading readiness of pre-schoolers is evaluated in terms of the job satisfaction of project staff members who train the mothers. While positive results in this area might be a desirable feature, they would not prove the program effective. Another type of inconsistency may occur in the statistical analysis of results, where *t*-test statistics are necessarily too high if the means and standard deviations presented in the same tables are assumed to be correct. Such inconsistencies are likely to provoke a generally skeptical attitude in panel members, or in any discerning reader.

Skepticism arising from the discovery of inconsistencies may often be removed by requesting further explanation of the

data. Such skepticism is always warranted, however, because there are several possible causes for the inconsistencies. Lack of agreement between numbers could be the result of a typographical error, or it could be an attempt to gloss over disappointing data. If fewer pupils were tested than were served, it could be the result of planned sampling that was part of the evaluation design, or of attrition that left a biased sample at posttest time. Errors in calculations may simply be mistakes, or they may be an attempt to make a "right" answer out of wrong data. Any errors, however, tend to detract from the overall credibility of the submission.

Are Enough Data Presented to Satisfy the Skeptical Evaluator?

Inconsistency is only one of the reasons why evidence may not be considered believable and interpretable. Lack of complete data also precludes accurate interpretation of an evaluation. Sometimes submissions omit important information such as the names, forms, and levels of tests used; the testing times; the number of students tested; or the number of students served. Data may be presented on only some of the measures that were administered. Once again, failure to report all of the data prompts the suspicion that those not reported were somehow unfavorable. Whatever the reason for the missing information, the result is that the evidence cannot be properly interpreted or taken at its face value.

In addition to basic information about the data that were collected, there should be a complete and accurate presentation of the analysis of these data. Types of scores should be clearly identified, e.g., raw scores, publishers' standard scores, Normal Curve Equivalents (NCEs), etc. Summary statistics should include both means and standard deviations. Each time scores are reported, sample sizes should also be reported. When statistical tests are used, they should be clearly identified; the rationale for their use, if not obvious, should be presented, and any assumptions made should be explained. Effectiveness can only be established by a complete and methodologically sound evaluation. Like inconsistencies in the data, missing or insufficient statistical information raises serious doubts about the adequacy of the evidence, and when judging the effectiveness of educational interventions, there is no room for "the benefit of the doubt." Where such doubt exists, responsible judges would have to take the more skeptical stance and withhold approval.

It should be mentioned that written submissions to the JDRP should be able to stand on their own merits. Although it has been common for evaluators to be present in order to answer some of the questions that may come up in connection with inconsistencies or gaps in the evidence, the panel at one point commended those evaluators who demonstrated enough confidence in their writeups to stay home. In this situation the importance of providing complete, consistent information becomes even more critical.

Are the Inferences Drawn from the Data Consistent with the Evidence?

Previous chapters have discussed some of the sources of mistaken inference. Two of the most common are the assumption of educational significance and the attribution of causation. Particularly if the advantages found for the participants in an intervention are statistically significant, the conclusion that they are educationally significant as well is all too often automatic. As we have seen, it is also all too often wrong. Similarly, when positive results appear, it is always easy and tempting to conclude that they were caused by the intervention, but often evaluators overlook such simple alternative causes as the fact that participants were a superior group to begin with.

As these examples indicate, the JDRP has to scrutinize evaluations carefully for evidence of mistaken inference, starting with errors as blatant as claiming that the numbers show something they do not, and including the much more complex judgments as well.

Has Evidence Been Presented that Common Errors Have Been Avoided?

Each evaluation design has its associated pitfalls and mistakes that can invalidate its results. The evaluator needs to be aware of these, and to set the reviewer's mind at rest by making it clear that they have been avoided. It is obviously important to provide a clear description of the evaluation design and methodology, and an essential part of this description is the indication that correct procedures have been followed at each step. For example, in describing an evaluation using a nonequivalent control group design,

the following sentence would not be very helpful: "The performance of the treatment group was compared with that of a control group from another school." Some assurance is needed that the control group was an appropriate one to use for the comparison. It should also be stated that the control group was chosen from another school nearly identical to the project school in terms of academic achievement, socioeconomic status, and regular educational program.

A major pitfall in some evaluation designs, particularly norm-referenced designs, is the regression effect error. It is important to specify clearly that the scores used for selection of participants were not the same as those used for measurement of pre-intervention status. They must not be the same. (Scores from an alternate form of the same test, however, are perfectly acceptable.) If participants were selected for a remedial reading project on the basis of their low scores on the XYZ Reading Test, and if those same scores were then used to figure the average pre-project status of the students, the gains attributable to the intervention would be overestimated. (For an additional discussion of the regression effect error, see Chapter IX.) Unless the evaluation report clearly states that different tests were used for selection and for pretesting, the reviewer has no assurance that this error was avoided.

Another important point to stress in the presentation of evaluation results is the objectivity of the data. Wherever it is possible for data to be contaminated, the write-up should describe measures taken to make sure that this did not occur. For example, a common source of problems is the procedure followed in testing. When tests are administered by persons with a stake in the outcome, either staff members of a project, or those with a close personal relationship to the subjects, those test results are suspect. In some intangible and perhaps totally unintentional way, the test administrators could have influenced the students' performance. They may have given extra directions, allowed more time, or deviated in some other way from the instructions for test administration. If the test required judgments or ratings by the administrator, their objectivity would be seriously in doubt.

To make it clear that there were no irregularities in testing, an evaluation should specify who gave the tests,

and under what conditions. For example, "Each participating class was given the XYZ Reading Test on May 1 (the same date that the national norm group was tested). The test was administered by teachers, who had been thoroughly trained in the publisher's instructions for proper use of the tests. Although it was not possible to obtain outside test administrators, the teachers were rotated so that they tested each other's classes, not their own." (This example scores another point, by the way, for making it clear that the error of testing outside the norming period was avoided.)

Another potential problem with impact assessments relates to the credibility of the measures used. Some types of measures are obviously more susceptible to spurious influences than others. For example, interventions designed to improve self concepts or attitudes toward school frequently rely on scales. A typical item might be, "How do you feel about being bused to school?"

Such scales are generally administered both before and after an intervention, and a higher number of smiling faces at posttest time is taken as a measure of success.

Attitude scales of this type are obviously transparent--that is, the respondent can tell exactly what attitude he or she is expressing. It is also unfortunately the case that respondents can always give the desired answer if they choose to do so--a situation that does not exist with most other types of tests. Because of this characteristic, drastically different scores can be obtained depending on the conditions under which the instrument is administered. Consider the following two examples: (a) "Children, as soon as we finish this test we will be getting on the bus for a trip to the zoo." and (b) "We're going to take this test today because it's raining and we won't be able to go to the zoo."

Attitude scales are not highly reliable or credible. Unless they have been very carefully devised so as to remove transparency, they probably do not constitute sufficient evidence of effectiveness. At the very least, it would have

to be convincingly demonstrated that the administration conditions were neutral both at pre- and at posttest time.

Much more convincing evidence could be obtained through unobtrusive measures such as absenteeism rates, number of books checked out of the library, number of books carried home at night, reduced vandalism, etc. Observations of classroom behavior can also be convincing if carried out by independent observers following structured procedures.

Questionnaires and opinionnaires are also often used in evaluating educational interventions. But findings such as "seventy-two percent of the parents polled considered the intervention 'moderately effective'" are difficult to interpret and unconvincing. Even if parents or other groups rated the quality of instruction higher after the intervention than they did before, the data are suspect for much the same reason as attitude scales. Respondents are likely to give some credit for effort even if they do not believe any real improvement was effected.

Other errors for which preventive measures should be described in the evaluation write-up might include "teaching to the test" (or other criterion measure); use of biased, invalid, or unreliable measures (particularly in the case of locally developed tests); improper selection procedures; and inadequate experimental controls on factors outside the intervention. In short, for every important error that is mentioned in this Ideabook, it is a good idea to provide evidence that steps were taken to avoid that error.

Examples of Convincing Evidence

The submission for the Math Labs for Disadvantaged Students in Honea Path, South Carolina, gave extensive evidence that errors had been avoided in the selection of students for treatment and control groups, possible "John Henry effects," regression effects, and contamination of test data. The following excerpts from the write-up are illustrative:

"A comparison group was selected from grades seven and eight in another school in the district. Both schools served communities having similar socioeconomic characteristics, offered similar educational programs, had similar educational

policies, and were alike in other relevant factors. The students in the control group were matched as closely as possible to project students on the basis of age, race, sex, and pre-achievement level on the Stanford Achievement Test. As an added precaution, the principal, the teachers and the students in the comparison school were not aware that a comparison group had been selected, and the selection test data for them were taken from the regular testing program of the district. The control subjects were tested in such a manner that they did not know that they were in any way involved in the project.

"Students selected to participate in the project scored below the 40th percentile on the Stanford Achievement Test, and their progress during the project was measured by the Comprehensive Tests of Basic Skills. This procedure minimized the confounding effects of regression to the mean on final test scores.

"The testing was under the direct supervision of the district guidance/testing coordinator, and the validation team was convinced that every possible precaution had been made to administer and score the tests in strict accordance with publisher's instructions.

"A matched comparison group which was unaware of being observed, the use of different instruments for selection of students and evaluation, the simultaneous testing of project and comparison students under identical conditions, and the close supervision of the testing without disrupting the usual school routine, assured that the data collected were as reliable as can be realistically expected in a public school setting. The controls and design employed also minimized threats to internal validity including regression to the mean, differential selection, history, and maturation."

Examples of Unconvincing Evidence

One project, designed to improve academic performance of low-achieving children, presented pre- and posttest scores for a ten-percent random sample of the students served. Unfortunately, this resulted in sample sizes sometimes as small as four, three, and even one! The panel concluded that such small N's rendered the data uninterpretable. The presentation of this partial information was particularly unusual

because the sampling was done after all students had been pre- and posttested, so that results for the total group could just as well have been used.

For another submission, panel minutes include the following comments about incomplete data and the resulting problems in trying to interpret evaluation results:

"Interpretability of the mean gains submitted depends on the belief the children were really in need to begin with. Lack of pre-scores made this interpretation difficult to sustain. Pre- and post-scores should be provided. More information on the selection procedures also was needed to permit interpretation. It was not clear if selection had been based primarily on pre-scores or on measures used for outcome criteria.

"The narrative style presentation of data was difficult to follow, and made certainty about the N's and data difficult. Data for FY 1975, which seemed critical in the discussion, should have been included. Presenting data differently for different years made interpretation unclear, particularly since so much comparable data, such as N's for the various mean gains, were missing."

For a number of projects, there was evidence that their data had been contaminated by use of test administrators whose ability to be objective was in doubt. In one case, in a remedial tutoring project, the standardized achievement tests used as pre/post measures were administered individually by each tutor to his or her own personal tutee-child. A similar problem was noted for an early childhood program in which each child was tested by a home visitor, who knew the children well.

Interestingly, the same two projects also shared another problem. Both had involved use of the pretest by the tutor or home visitor to diagnose each child's needs and to design instruction accordingly. Thus throughout the year, the project staff members were actually "teaching to the test."

IX. COMMON EVALUATION HAZARDS AND HOW TO AVOID THEM

Each of the following subsections provides a brief description of a common evaluation hazard and what should be done to avoid it. The hazards are presented in the approximate order of their judged frequency of occurrence and severity. Not all hazards, however, are relevant to all evaluation designs. An attempt has therefore been made to describe as early as possible in each presentation the situations where each hazard is likely to arise. Unfortunately, because many of the hazards are relevant to two or more designs, it was not feasible to organize this chapter according to evaluation design.

Hazard 1: Claiming Much, Providing Evidence of Little

The goals of an innovation may be multiple and grand; the evidence may be modest and limited, as in one instance where preventing school failure through increased parent involvement, home/school cooperation, building positive attitudes, and academic gains were claimed while the only data were pre-post measures of spelling.

In another submission, it was claimed that participants would get in less trouble with the law. Evidence for this claim was not included in the submission; in a discussion, the claim turned out to be based on the fact that "The children are being watched from 8:30 to 3:30 every day. This would help keep them out of trouble." In the same submission, the claim that regular progress during the fall-spring year resulted from the summer program turned out to mean "If our children gained three months in six weeks, they'd be gaining the equivalent of one year in eight months."

The panel can only reach judgments based on the evidence; where evidence matches the claims, a favorable decision is far more likely than where evidence falls far short of goals, objectives, and claims.

Hazard 2: Selecting Measures Not Logically Related to the Intervention

Valuable as standardized achievement tests may be for assessing many interventions, in many instances there are other, more sensitive, more appropriate sources of data. Measures should be selected because they will assess the central consequence of the intervention, be it improved career decision making, adequate knowledge of human reproduction or contraception, improved consumer skills, minimum competencies for self care in dressing and toiletting, increased pride and self confidence as a member of one's ethnic or racial group, improved teacher classroom control, or a better school planning and budgeting system.

For almost all educational outcomes, reliable sensitive measurement is possible, using techniques as varied as analysis of basic school records, interviews, questionnaires, case studies, observations, and tests.

Hazard 3: The Use of Grade-Equivalent Scores

Grade-equivalent scores present a severe methodological problem when used in norm-referenced evaluations. No norm-referenced evaluation based on grade-equivalent gains can be accepted. When used with control-group or regression-evaluation models, the methodological problems are far less severe although problems of interpolation remain.

Grade-equivalent scores provide an insensitive, and, in some instances, a systematically distorted, assessment of cognitive growth. The concept of a grade-equivalent score is misleading--for example, a grade-equivalent score of five attained by a third grader on a math test does not mean that he knows fifth-grade math. Possibly he can do third-grade math as well as the average fifth grader, but it is likely that no fifth-grade students have ever taken the third-grade level of the test.

The use of grade equivalents for evaluation purposes creates a second problem in that they do not form an equal-interval scale and should never be averaged. Finally, grade equivalents are constructed based on the assumption that growth occurs at the same rate throughout the school year.

Research has shown, however, that learning typically does not follow this regular pattern and, consequently, gains measured in grade equivalents will be artificially inflated or reduced.

To avoid this hazard, it is recommended that normalized standard scores (often provided by test publishers) be used for computations and that gains be interpreted through reference to pre- and posttest percentile standings with respect to a norm group.

Hazard 4: The Use of a Single Set of Test Scores for Both Selecting and Pretesting Participants

When students are selected for participation in a special group because they obtained relatively high or relatively low scores on some test, use of these scores as pretest measures invalidates any kind of norm-referenced evaluation. It can also invalidate control-group evaluations when the treatment and control groups are drawn from different populations (see Hazard 8). This problem stems from what is known as "statistical regression," "regression toward the mean," or simply, the regression effect.

If students are selected from a group because of their low test scores and are retested on the same or a comparable test, they will tend to score higher on the second testing, while an initially high-scoring group will score lower. The result is that low-scoring groups appear to learn more from a special program than they actually do, while gains in special programs for high-scoring students may be obscured.

To avoid this hazard, students should be selected for participation in a special treatment based on one set of test scores and then be pretested using an alternate form of the same test or a different test. Another alternative is to base student selection on some form of needs assessment other than test scores.

Hazard 5: The Use of Comparisons with Inappropriate Test Dates for Obtaining Information

In norm-referenced evaluations, tests should be administered at nearly the same time as the test publisher tested

the norm group. When control groups are available, few evaluators would consider testing the treatment and control groups more than a few days apart. When norms are used as a substitute control group, this same consideration needs to be given to test dates.

To avoid this hazard, treatment group students should be tested within two weeks of the midpoint of the interval during which the normative data were collected. Particularly where the evaluation spans a summer, both pre- and posttest should deviate from the norming date in the same direction and by the same amount of time. Testing within six weeks of empirical normative data points is permissible if linear interpolations or extrapolations of the normative data are made. Tests that provide normative data for only one point in the year should not be used in fall-to-spring norm-referenced evaluations. The principles of comparability in time of data collection also apply to other outcomes, such as job placement rates that will show systematic seasonal variation.

Hazard 6: The Use of Inappropriate Levels of Tests

If most of the pupils achieve very high or very low test scores, the level of the test may be inappropriate for assessing their performance. If pupils encounter the test floor at pretest time or the ceiling at posttest time, treatment effects will be underestimated. Conversely, if the ceiling is encountered on the pretest or the floor on the posttest, gains will be overestimated. Ideally, students should score in the middle of the range of possible raw scores. Average performance for groups of students should generally fall between 30 and 75 percent of the total number of items answered correctly.

To avoid this hazard, test levels should be selected on the basis of the achievement levels of the students, not on the basis of their grade in school. In most cases, the nominally recommended test level or one level above or below will be suitable for testing most groups of students.

Using a test level other than that nominally recommended for a particular grade is likely to mean that norms tables for the tested students are not included in the test manual for that level. However, it is not meaningful to assess

either status or gains by comparisons with students at a different grade level. The status of a sixth grader should be assessed using sixth-grade norms even if he is tested with a fourth-grade test. Most major test publishers, fortunately, have interlocked their test levels by providing an expanded standard score scale which enables the determination of score equivalencies between adjacent test levels. These scores make it possible to predict from a pupil's score on one test level how he would have scored on the next higher or lower level, thus providing access to the in-level norms. Another possibility (although rare) is that the manual may provide normative data for pupils tested out of level.

Hazard 7: Missing Data

Regardless of the evaluation model used, data analysis should be based only on those students with both pre-treatment and post-treatment scores. Interpretation of these data, however, should take into account the characteristics of the students who dropped out, entered late, or graduated from the project. For example, if all of the lowest scoring students on a pretest dropped out before posttest time, the average posttest score would increase with respect to the pretest scores simply because of the missing students. This increase could be misinterpreted as a gain. Likewise, if the high-scoring students graduated from the group, the mean posttest score would be artificially deflated.

To avoid this hazard, every effort must be made to obtain pre-treatment and post-treatment data for each project participant, and to base comparisons on those students for whom both data elements are available. Data from students having only pretest or only posttest scores must be carefully examined to see if they differ in some systematic way from the data of students having both pre- and posttest scores. A description of any of these differences should be included in the evaluation report.

Hazard 8: The Use of Noncomparable Treatment and Control Groups

In "true" experimental designs, treatment and control groups should be similar in all educationally relevant

respects before the treatment begins. Groups that differ in terms of pretest scores present an obvious source of bias. Other more subtle factors such as differences in age, sex, race, or socioeconomic status can also exert strong biasing influences and should be avoided. Nonvolunteers should never be used as controls for pupils who volunteered (or were volunteered by their parents) for a particular instructional treatment.

Whenever possible, students should be assigned to treatment and control groups on a random basis. For example, with a semester-long reading program, pupils could be randomly assigned to first- or second-semester groups. For the first half of the year, one group would serve as the control group for the other, but both groups would ultimately receive equal amounts of the treatment.

In some cases, pre-existing groups will be enough alike so that they can appropriately be considered equivalent to random samples from a single population. In other cases, a control group will be known to differ systematically from the treatment group. Where the difference is small, the control group model may still provide the best method of evaluating the project, and statistical adjustments can be made to compensate for between-group differences. Where the differences are large, however, there is no way in which a noncomparable control group can provide an accurate estimate of how well the treatment group would have done without the treatment.

Hazard 9: The Use of Inappropriate Statistical Adjustments with Nonequivalent Control Groups

There are several statistical procedures that are widely used in an attempt to compensate for initial differences between treatment and control groups. Some are legitimate while others are not. Making between-group comparisons using either "raw" gain scores or "residual" gain scores falls into the latter category. Both procedures should be scrupulously avoided.

A raw gain score is simply the difference between a pre- and a posttest score and reflects the gain made between testings. It is argued that, although two groups may have been

somewhat different in terms of initial achievement levels, their expected gains would be roughly comparable after the same educational treatment. This would be true, however, only when each group's posttest standard deviation is the same as its pretest standard deviation. Where the posttest standard deviations are larger than those of the pretest scores, a raw gain score analysis will systematically underestimate treatment effects. Conversely, the procedure will systematically overestimate treatment effects where the standard deviations of pretest scores exceed those of posttest scores.

A residual gain score is the difference between an actual posttest score and a posttest score estimate derived from the combined treatment and control group regression line. Presumably the mean residual gain score for a group that received an effective treatment would be positive while that for the control group would be negative. Also, the sum of the absolute values of the two differences would provide an index of the size of the treatment effect. Unfortunately, it can be shown alegbraically that a residual gain score analysis always underestimates the size of the treatment effect except where the groups' pretest scores are equal. Furthermore, the amount of underestimation is directly proportional to the size of the initial difference between groups.

There are other factors, such as how the treatment and control groups were formed, which determine the appropriate adjustment procedure to compensate for their initial differences. The issues involved, however, are very complex. Expert advice should be sought.

Hazard 10: Constructing a Matched Control Group after the Treatment Group Has Been Selected

Finding "matches" for treatment participants in some other group is a fundamentally unsound practice. Unless they and the treatment pupils are equally representative of the groups from which they are drawn, statistical regression will act differentially on the two groups and artificially inflate the apparent gains of one group with respect to the other.

In the most common situation, the group(s) from which the matching control pupils are drawn will be higher achieving than those from which the treatment group pupils are selected. Consequently, the control group pupils will be farther below the mean of the group(s) to which they belong than the treatment group children. On retesting they will thus show greater statistical regression and their posttest scores will be too high to serve as a no-treatment expection for the treatment group participants.

The correct procedure for establishing matched control groups is to do the matching first and then assign members of each pair randomly to the treatment or the control group. That is, a large group of students, all eligible to be in the project, must be available. The first step is to divide the group into matched pairs based on test scores, ethnic background, sex, etc., so that the two members of each pair are as similar as possible. Then, after the matching process is complete, some random procedure such as flipping a coin should be used to decide which member of each pair goes into the treatment and which into the control group.

Since this procedure will rarely, if ever, be possible in real-world situations, the only way to avoid this hazard is to avoid matching. A better practice is to work with unmatched and therefore slightly different groups and to correct for between-group differences with appropriate statistical adjustments, such as some form of covariance analysis. Again, practitioners should seek the help of experts.

Hazard 11: The Careless Collection of Data

Testing and other types of data collection must be accomplished with scrupulous attention to detail. For most evaluation models, the primary requirement is that treatment and control or comparison groups be treated in exactly the same way. Minor variations from the procedures described by the test publisher are permissible in control group models and in certain quasi-experimental designs. It is essential, however, that treatment and comparison groups be tested in exactly the same manner. In norm-referenced evaluations, treatment group students should be tested in the same way as the students in the norm group. This requirement means that procedures outlined by the test publisher must be followed precisely.

Problems arise if tests are administered or scored in an inconsistent and careless manner. If there are differences in the ways in which the test takers and the norm group students are tested or if there are differences in the procedures, conditions, and scoring at pretest and posttest times, then it is impossible for the resulting data to be accurately interpreted. There are no statistical manipulations that can compensate for mistakes made in administering or scoring a test.

To avoid this hazard, the following steps should be taken:

1. Test procedures must be orderly and accurate if scores are to be meaningful.

2. Test administration and scoring procedures must be exactly the same for the treatment group as for the control, comparison, or norm group used to generate the no-treatment expectation. Testing treatment group pupils in exactly the same way as pupils in the norming sample means following the test publisher's directions in every detail.

3. The procedures, conditions, and scoring methods used during posttesting must be exactly the same as those used during pretesting.

Hazard 12: The Use of Different Instruments for Pretesting and Posttesting

In the norm-referenced design, it is not advisable to change tests between pre- and posttesting because there is no adequate way to compare pretest scores on one test with posttest scores on a completely different test. Since each test publisher follows slightly different norming practices, it is likely that one test's norms will be slightly "easier" than another's. These differences do not matter if the same test is used both pre and post but could magnify or obscure real gains if changes were made. While it is not essential to use the same form and level of an achievement test pre and post, this practice is also recommended.

Some tests have been developed so that the lower levels are intended for use at the end of one grade and the beginning of the next. In these instances, to use the same form and level of test for fall pretesting and spring posttesting, it will be necessary either to pretest or posttest out of level. In some grades where spring-to-spring or fall-to-fall evaluations are conducted, it may be necessary to change test levels in order to avoid ceiling or floor effects; unfortunately, this practice will introduce an unknown amount of error into the measure of gain.

Hazard 13: The Use of Inappropriate Formulas to Generate No-Treatment Expectations

Many projects use an unrealistic theoretical model or formula to calculate "expected" posttest scores from IQ or other pretest scores. If students do better than the calculated expectation, the project is considered a success.

Many methods have been devised for calculating performance-level expectations that rest on untenable assumptions. Neither IQ scores nor grade-equivalent scores should be used to generate no-treatment expectations. For example, a student who has gained .7 year per year, on the average, since beginning school, is presumed to continue at the same rate unless a special program increases his rate. Unfortunately, grade-equivalent gains measured from fall to spring will usually exceed this rate--even for typical Title I children--and treatments will appear to be more effective than they really are.

In norm-referenced models, this hazard may be avoided by generating the no-treatment expectations solely from empirical percentile norms tables. When control groups are used, the actual posttest scores of these groups provide the proper basis for evaluating treatment effects.

Hazard 14: Mistaken Attribution of Causality

Observed gains may have resulted from the Title I treatment, but there are always plausible alternative explanations. The plausibility of these alternative explanations should be carefully examined before evaluation results are attributed

to project impact, as evaluation hazards are often the cause of apparent gains or losses.

Sometimes project participants learn substantially more than would have been expected, but the project, per se, is not responsible. Instead, the gains could be a result of the Hawthorne effect in which special project participants do well simply because they are getting special attention. The nature of the treatment may not necessarily be important. An opposite result may follow from a John Henry effect (Saretsky, 1972). In this case, comparison group students work extra hard to prove that they are just as good as project students.

Other likely causes of misleading gains are unrecognized "treatments" which have nothing to do with the project. Most school systems are in a constant state of flux with multiple changes every year. Changes in school programs, personnel, facilities, class sizes, community characteristics--any or all of these factors can affect student performance. Also, the true source of achievement gains is sometimes improperly identified because children are involved in more than one treatment. Under these conditions, it is impossible to determine causality in an unambiguous manner.

A Final Thought

This entire document has been filled with cautions, descriptions of hazards, lists of practices to avoid, and examples of unsatisfactory evaluation and reporting procedures. Taken together, these discussions may have given the reader the impression that JDRP approval is very difficult to obtain, that standards are unrealistically high, and that only sophisticated professional evaluators will be able to prepare acceptable submissions. Such inferences, however, are completely unwarranted.

Evaluation is not a particularly mysterious process and, if educational interventions are truly beneficial, it should be a relatively straightforward matter to assemble credible evidence of success and to present it in an acceptable manner. Many interventions have been approved by the JDRP, and the majority of the approved submissions were prepared by local personnel who were not professional evaluators.

Use of common sense in preparing a case that (a) something has happened, (b) that it has been observed often enough to be credible, (c) that it is educationally important, (d) that it can be duplicated in other settings, (e) that it was the result of the intervention and not some other influence, and (f) that the evidence supporting these claims is convincing will generally be sufficient to obtain JDRP approval. This Ideabook was not designed to discourage potential submitters but rather to provide constructive guidance and thereby enhance the probabilty that effective educational interventions will be identified, approved, and disseminated as quickly and efficiently as possible.

X. FORMAT AND INSTRUCTIONS FOR SUBMITTING MATERIALS TO THE JOINT DISSEMINATION REVIEW PANEL

The outline on the following pages displays the format that should be used for submitting materials to the Joint Dissemination Review Panel. Under each major heading are specifications for the kinds of information that should be included.

In order to be convincing, the evaluation need not be a strict experimental design, although this type of evidence is desirable. However, there should be some kind of high-quality, objective, methodologically sound, quantitative evidence demonstrating that the intervention in question was effective and superior to other, more commonly used methods or approaches, and that the observed effects were caused by the intervention.

The total length of each submission, including all descriptive material, tables, etc., should not exceed ten pages. Ten pages is a maximum; less lengthy presentations are definitely acceptable and encouraged. The evidence needs to be convincing, not necessarily lengthy. Some of the strongest arguments are brief, concise, and to the point.

While brevity is a virtue, the materials should include all the information the panel will need to make its decision. All submissions should follow the format by using the headings shown, and should cover all the points mentioned in the instructions under each heading. The points under each heading should be presented in the order most convenient and logical for the intervention being documented, not necessarily in the same order as in these instructions.

FORMAT FOR SUBMITTING MATERIALS
TO THE JOINT DISSEMINATION REVIEW PANEL

PROGRAM AREA: (e.g., Title III, reading, career education, environmental education, education for the handicapped)

I. INTERVENTION TITLE, LOCATION:

Specify the title of the intervention and the location for which evidence of effectiveness is being submitted.

II. DEVELOPED BY:

Indicate who developed the intervention originally, even if this happened at a different site than the one for which evidence of effectiveness is being presented.

III. SOURCE AND LEVEL OF FUNDING:

List all funding sources for the intervention at the location for which evidence of effectiveness is presented and, for each source, list the amount of funds (see Figure 1 for an example).

IV. YEARS OF INTERVENTION DEVELOPMENT:

Indicate the year or years during which the intervention was originally developed or tested.

V. BRIEF DESCRIPTION OF INTERVENTION:

Briefly describe the intervention for which claims of effectiveness are being made. The description should cover at least the following points:

What is the intervention?
What are its objectives?
What claims of effectiveness are being made?
What is the context in which it operates?
Who are the intended users and beneficiaries?
What are the characteristics of the groups on which the intervention was developed and tested?

What are the salient features of the intervention?
What are the costs for adoption and maintenance of the intervention? (See table on next page.)

VI. EVIDENCE OF EFFECTIVENESS:

Describe the evidence of effectiveness for the intervention. This section should deal with each of the following points, although not necessarily in the same order:

Interpretability of measures: Evidence that the quantitative measures are reliable and valid indicators of the effects claimed.

Credibility of evidence: Who collected and analyzed the data, what assurances are there that the findings are objective?

Evidence of impact: What is the evidence that something happened? What are the effects claimed for the intervention?

Evidence of statistical reliability of the effects: What is the evidence that the effects happened often enough and with sufficient reliability to be likely to happen again under similar circumstances?

Evidence that the effects are educationally meaningful: What is the evidence that the effects are large enough, powerful enough, or important enough to be educationally meaningful, regardless of their statistical significance?

Evidence that the effects are attributable to the intervention: Can alternative explanations such as practice effects, maturation, selection of superior treatment groups, etc., be ruled out?

Evidence of generalizability to the populations for which the product or practice is intended: Evidence that the product or practice has been tested widely enough and under sufficiently diverse circumstances to give assurance that the effects claimed may be similar when the product or practice is used elsewhere for the populations intended.

Source and Level of Funding of Intervention		
	INSTALLATION (Non-recurring Costs)	SUBSEQUENT YEARS (Recurring Costs)
Personnel		
Personnel Training		
Facilities		
Equipment & Materials		
Consumables		
Other Costs* - Specify:		
_____		
_____		
_____		
TOTAL		

*E.g., transportation, technical assistance, public relations, etc.

Figure 1. An illustration of a table shell for showing costs.

XI. A MODEL SUBMISSION

PROGRAM AREA: Disadvantaged Children (Title I) Reading

PROJECT TITLE, LOCATION: Cycle 10 Reading Project
Mediopolis, Florifornia

DEVELOPED BY: Reading Specialists, Mediopolis Public Schools

SOURCE AND LEVEL OF FUNDING: ESEA Title I $155,804

YEARS OF PROJECT DEVELOPMENT: 1975-76

BRIEF DESCRIPTION OF PROJECT: Cycle 10 is a pullout program that provides intensive remedial reading instruction for disadvantaged students at the elementary level.

Goals and Objectives

A. Seventy-five percent of the participants will improve their percentile standing in reading with respect to national norms.

B. Cycle 10 pupils will increase their independent reading during the period of instruction.

C. Pupils will be able to relate the sound to the symbol of all letters with an accuracy of 70%.

D. Pupils will be able to identify rhyming words with an accuracy of 80%.

E. Pupils will be able to read orally with reasonable fluency and answer comprehension questions at their level.

F. Pupils will read a good variety of reading materials and show gains in vocabulary development.

The evaluation on which claims of effectiveness are based focused upon the first objective, improvement in reading achievement.

Context and Students Served

Mediopolis, Florifornia, is an urban community covering 17.4 square miles with a metropolitan population of approximately 800,000. The community is about 46% Black, 38% Anglo, and 16% Puerto Rican. English is a second language to many Puerto Ricans in the community, and a nonstandard dialect is prevalent among the Blacks.

The Cycle 10 program serves children from six "high need" schools (schools where the mean annual income per family is no more than $4000). Students attending these schools live in neighborhoods characterized by extensive public housing projects and steadily deteriorating, crowded living conditions. In 1969, nearly half of the children came from families receiving public assistance and 70% qualified for state and city financial aid. A large segment of the parents who were employed had few marketable skills, with many holding unskilled and semi-skilled jobs.

The schools range in enrollment from approximately 300 to 1,800. Most encompass kindergarten through grade eight, but some run only through grade four and others through grade six. The average class size is 25, and the student-teacher ratio is 25 to one. The average IQ of these students as measured by the Lorge-Thorndike Test is about 90. The average per-pupil expenditure for regular school programs was $856 during the 1968-1969 school year.

The Cycle 10 program is focused primarily on third-grade children but also serves some fourth- and a few second-grade children. There are three Title I-funded centers, each of which operates in three ten-week cycles per year. Forty-five children are served at each center during each cycle.

Program Description

A Cycle 10 center includes three teachers each having a separate, open-space classroom and each specializing in one of the following three reading areas: (a) decoding, (b) vocabulary and comprehension, and (c) individualized reading. The 45 children are divided into

groups of 15 at the beginning of a cycle. Each group spends an hour with each teacher every day. The Cycle 10 instruction consumes the entire morning of the school day and the program teachers spend the afternoons (a) preparing lessons and keeping records for each of the 45 children, (b) making new instructional materials (c) coordinating with classroom teachers, and (d) participating in professional development and training sessions.

The 45 children for a single cycle are all chosen from a single school. For one or two of the cycles, children come from the schools in which the centers are located. For the remaining cycles, children are bused from other Title I schools. Children are nominated for participation by their regular classroom teachers with the final selection being made by the Cycle 10 staff. The criteria for nomination/selection include the following: (a) children should be reading both below grade level and below expectation, (b) children should not be in ESL, Bilingual, or IIC programs, and (c) children should have good attendance records and be able to work successfully and cooperatively in an intensive type of program.

Each center is staffed by a team leader who is a certified reading specialist, two team teachers who are certified reading teachers, and a half-time secretary. In addition, there is a half-time project director who has overall responsibility for the three centers. The project director participates actively in all operations, visiting each team at least weekly. This person organizes and participates in training sessions and exercises final authority on teaching methods and materials.

Each Cycle 10 team has considerable autonomy. Members are expected to exercise skill and imagination in developing teaching methods, producing original materials, and selecting commercial materials. Each team controls a budget for purchasing equipment and materials.

Cycle 10 uses a flexible, largely individualized approach employing a wide variety of commercial materials supplemented by many teacher-made materials, displays, and modifications to commercial materials. Guidelines

for each of the three reading instructional areas have been developed and are provided to new Cycle 10 teachers, but the primary responsibility for choosing appropriate instructional strategies rests with the individual teachers. Specific provisions have been made so that each teacher receives inputs from (a) the project director through informal visits and formal training sessions, (b) other members of the Cycle 10 center team through daily contact, (c) members of the other Cycle 10 teams through regularly scheduled meetings, (d) the children's regular classroom teachers through meetings and written communications, (e) reading consultants from the sending schools through regular meetings, and (f) parents through informal visits and irregularly scheduled gatherings.

Costs

The 1972-1973 budget for three Cycle 10 centers was $155,804. These three centers served 405 pupils from six schools for an average cost of $385 per pupil. Costs for salaries included three reading consultants, six teachers, and two secretaries, plus part of the project director's salary. The school district provided three classrooms and a small office space for each center. The following table presents costs for project installation and for subsequent years.

Sources and Level of Funding for Cycle 10

	INSTALLATION (Non-recurring Costs)	SUBSEQUENT YEARS (Recurring Costs)
Personnel	$134,604	$134,604
Personnel Training	5,272	--
Facilities	--	--
Equipment & Materials	13,662	2,835
Consumables	5,087	1,665
Other Costs - Specify:		
transportation	15,200	15,200
public relations	3,390	1,500
TOTAL	$177,215	$155,804

EVIDENCE OF EFFECTIVENESS

Cycle 10 teachers administer several standardized tests on a regular pretest-posttest basis. Of these, the Cycle 10 Evaluation Report prepared by the evaluation office of the Mediopolis Public Schools presents mean pre- and post-test scores from the 1957 edition of the California Achievement Test (a well known battery with high reported reliabilities) as evidence of program effectiveness. Mean total reading scores for 1971-1972 Cycle 10 pupils were raised from 2.7 to 3.3, a gain of six months for ten weeks of instruction. In 1972-1973, pupils went from 2.3 to 3.2, a gain of nine months in ten weeks.

The interpretation of these results may be open to some question due to the short time between pre- and post-testing and the inherent difficulties in interpreting grade-equivalent scores. Therefore, a more conventional analysis

was undertaken by the present evaluators using data from a city-wide testing program. Under this program all third graders were administered the CAT (1970 edition) in September of 1972 and again in May of 1973. Tests were administered by the regular classroom teachers, not Cycle 10 teachers, under the supervision of personnel from the district's central office. The test scores were available in the form of computer printouts, and it was possible to find pre- and posttest scores for 134 of the Cycle 10 third graders from four of six participating schools. Using Cycle 10 test scores, these 134 pupils were compared with the Cycle 10 pupils for whom city-wide test scores were missing from the computer printouts. There was no evidence to suggest that the 134 pupils differed substantially from the remaining Cycle 10 pupils.

Initially, the non-Cycle 10 pupils were considered as a possible comparison group, and their scores from the city-wide testing program were examined. However, examination of their scores and discussions with Cycle 10 personnel led to the conclusion that the non-Cycle 10 pupils were inappropriate as a control group for two reasons. First, they were not, in general, similar to the Cycle 10 pupils, being divided between the above-average learners who did not need special reading instruction and the very slow learners who were below the minimum Cycle 10 requirements. Second, they were not treated as Cycle 10 pupils would have been treated in the absence of any special reading program. In fact, some non-participating pupils received more individualized attention than the Cycle 10 pupils especially in cases where most of the class spent ten weeks in Cycle 10, leaving the regular classroom teachers free to work with a small number of pupils.

In lieu of an appropriate comparison group, Cycle 10 scores on the 1970 CAT were analyzed in relation to the national norms. Raw scores of individual students were used to compute mean percentile levels for each of the four schools having adequate data. These mean scores and the means over all four schools are given in the table below.

MEAN CAT (1970) SCORES FOR CYCLE 10 THIRD-GRADE PUPILS

		September 1972		May 1973		Gain	
School	N	Raw	Percentile	Raw	Percentile	in S.D. Units	t
Lewis	42	32	12	51	24	.47	3.00**
Battery	28	28	7.5	39	12	.26	1.37
Grape	39	29	8	57	28	.82	5.07**
Washington	25	28	7.5	44	16	.45	2.18*
Total Group	134	29	8	48	22	.63	7.30**

*Significant at .05 level, two-tailed test.
**Significant at .001 level, two-tailed test.

As this table shows, growth in relation to national norms was statistically significant for students in three of the four schools, and for the average of the total group. More important in terms of educational significance, all the gains were on the order of one-third of a standard deviation or better, with an average total group gain of .6 of a standard deviation in relation to national norms.

While the relatively small proportion of fourth graders in the program was not included in this analysis, an analysis of twelve-month gains for first-cycle fourth graders reported in the Cycle 10 evaluation suggests that they gain even more than the third graders.

Cycle 10 has not, to our knowledge, been replicated at any other locations. However, its successful operation in more than one school, continuing over a period of eight years, provides a strong indication that the project would be widely generalizable to other disadvantaged populations.

XII. ANOTHER MODEL SUBMISSION

PROGRAM AREAS: Career Education, Drop-out Prevention, Recovery

I. Title of Program: The Career Intern Program (CIP)

II. Developed by: The Opportunities Industrialization Centers of America, Inc. (OICs/A), a private, minority-operated, national education and training organization. The development site for the CIP program is Philadelphia, Pennsylvania.

III. Source and Level of Funding:

a) Development and testing period: Initial award (1971) from the U. S. Office of Education, and subsequent awards from the National Institute of Education. Total Federal funds for CIP development, testing and evaluation: $2,919,563.

b) Funds for continuing operation: Philadelphia School District and the State of Pennsylvania Department of Education. Total non-Federal funds for CIP operation in Philadelphia: $291,666.

IV. Years During Which the Program Was Developed and Tested:

The development and testing period was 1972 through 1975.

V. Brief Description of the Program:

A) What is the Career Intern Program?: The CIP's long-term goals are to help minority and low income youths achieve success in higher education, work and their personal lives. The short-term goals are to help young people between the ages of 16 and 21, who are at risk of dropping out of school or who have already dropped out (1) to complete high school, (2) to acquire occupational knowledge and career planning skills, (3) to gain basic skills, and (4) to make the transition from high school to further education or to work.

Career education's philosophy has been selected as the organizing approach because it is believed that learning how they can plan for their own career development will be a powerful motivating force for these severely at-risk youth. The motivation given by career education and the OICs/A ethos is coupled with a strong, focused, individualized educational program provided in a separate alternative high school. The school is operated by the OICs/A as the Career Intern Program in the Germantown section of Philadelphia.

Germantown is a predominantly Black, economically depressed area beset with gangs, unemployment, drugs, and high drop-out rates, among other problems. All the youth participating in the program were a year or more behind their age mates in Carnegie credits and/or had left school.

With regard to the program itself, drop-outs and potential drop-outs from traditional 10th, or 11th or 12th grades entered a school which combines career investigation, counseling and basic academic training with individual attention, support, and community and parent involvement. The program is an outgrowth of OICs/A experience in serving adult men and women needing skill training (DOL/CETA awards).

Recruitment of interns (all voluntary) includes self-referrals in response to community outreach, and referrals from high school counselors. Applicants are interviewed and tested; applicants with severe drug problems or with extensive prison records are screened out. During development and testing, selection of applicants was done on a lottery basis.

Once selected, interns enter Phase I, a 21 week career awareness program. This phase includes classes in English, Math, Social Studies and Science infused with career education. Instruction is keyed to the working world, showing interns how each discipline relates to occupations. A low (10:1) intern/staff ratio assures individual attention, backed up by weekly intern/counselor meetings and career counseling seminars.

Phase II provides up to four semesters of hands-on, fused academic and career education, including two week-long placements in the field, aimed at testing the interns' interests and providing a realistic perspective of the demands and rewards of jobs they explore. In addition to hands-on exposure, Phase II stresses individualized instruction and independent study. Advanced courses in the disciplines mentioned above are required.

The length of time necessary for any given student to complete Phases I and II depends upon the number of credits earned prior to entry into CIP. The number of months for completion described above are "outer limits". The average time necessary to complete the program is about 10 months.

In Phase III, interns concentrate on making the transition from school to work, or to advanced education. Interns who choose vocational or on-the-job training are assisted by their counselors for six months. College-bound interns can be assisted by their counselors for a full year after leaving CIP. "Success" is defined as (1) completing a regular high school diploma or GED or (2) remaining in school and working toward these.

B. Claims of Effectiveness: When operated under OICs/A guidance for Black urban youths, the following is claimed:

- Relative to true controls, CIP participation will be associated with an increase in school graduates, increased retention and decreased drop-outs.
- Relative to true controls, CIP participation will be associated with improved career planning, information and decision-making skills, and with increased utilization of career resources.
- Relative to true controls, CIP participation will be associated with improved reading and mathematics skills.
- Relative to true controls, CIP participation will be associated with improved transition from high school to post-secondary education or work.

It is further claimed that the improvements will be educationally meaningful in magnitude, and within reason with regard to cost.

C. Intended beneficiaries: Although intended for all minority and low income youth at the risk of dropping out of school for reasons apparently associated with poor motivation, CIP has been tested only for Black, urban youth. Claims of effectiveness are made only for mostly Black, urban populations.

It may be useful, however, to have a picture of what the CIP applicants looked like at time of enrollment in the program. Most of the students who enter have a history school-related problems which indicate that the likelihood of their graduation from high school is extremely low. Almost half of the students applying to CIP had already dropped out of school, and the remainder had been classified by their counselors as potential drop-outs. The majority of applicants were at least one year behind their peers in the number of school credits earned. (Source: intake interviews and previous school records obtained from counselors.)

Further, most of the students at time of application to CIP were four years behind most students in the United States in reading and mathematics achievement (Source: Stanford Achievement Test Advanced High School Battery, intake testing).

While most students had a very negative view of themselves as students, they demonstrated, on a non-verbal test of reasoning ability, average levels of ability to learn. Also, their general self-esteem was high, and they saw themselves as people who liked and were liked by others. They reported their relationships with their families as characterized by love and mutual respect. (Sources: Coopersmith Self-Esteem Inventory and Raven's Stanford Progressive Matrices Test, intake testing).

With regard to career information and career planning ability, applicants had little factual knowledge about careers but had previously engaged in actual exploration of careers and planning. (Source: Super's Career Development Inventory, intake testing).

Further, while most applicants were oriented toward vocational and technical areas in their career aspirations, their parents hoped they would enter professional careers which required advanced education. With regard to the life experiences of students applying to CIP, most applicants came from families in which neither parent had graduated from high school. In addition, almost all applicants came from neighborhoods in which gang activity and drug use were epidemic. A majority of applicants came from families in which the primary wage earner was a woman, and most of the applicants indicated that they had to work part-time to supplement the family income. Finally most of the applicants came from homes in which the primary breadwinner was engaged in a low-status occupation. (Source: intake interviews with students and parents).

D. Costs: See section VII for cost table sheet. The per student costs during start up and program operations are reported in Table III. The cost of materials development and evaluation have been excluded. It is assumed that school space would be donated by the District. On this basis, the per pupil cost for the operational years during which the program was summatively tested was $2,384.

For comparison that the cost per student in Philadelphia is $1,615 for a regular high school; $2,314 for a vocational high school, and, for an exclusive private school, the per student cost is $2,784. CIP costs more than a regular high school, about the same as a vocational high school, and less than a private school.

E. Interpretability of the Measures: Data collection combined standard, commercially available tests, ethonographic observations, individual case studies, unobtrusive measures, and on-going observations of the program. All commercially available tests are described in Buros' Mental Measurements Yearbook, and all have been used elsewhere for measuring the objectives for which they are reported. These are the Stanford Reading and Mathematics Achievement Tests, Super's Career Development Inventory, and Coopersmith's Self-Esteem Inventory.

F. Credibility of Evidence: The research design involved three separate cohorts of applicants, applying at six to eight months intervals. Through over-recruitment and a lottery process, known in advance to all applicants, three separate sets of true experimental and control students were selected in a ratio of about 3 interns to 1 control. The ratio was selected to permit maximum entry into CIP with the minimum N estimated to be needed for a control group large enough to be sensitive to educationally meaningful effects.

All tests were administered by employees of the third party evaluator (Gibboney Associates). Employees did not know which persons were interns and which controls. All other data (e.g. high school completion) were collected by the third party evaluators. In addition, all data and analyses have been audited by Dr. Andrew Porter of Michigan State University. In analyzing the data, three different statistical techniques were used, all of which yielded the same results. The data presented at this submission were analyzed through application of the co-variance model, using pre-tests as covariates. Data used in this analysis focused exclusively on those students who completed all instruments both pre and post--a smaller N than analyses using all available data.

VI. Evidence of Effectiveness:

A. Evidence of Reliable, Meaningful Effects

TABLE I: School retention (data pooled over three cohorts)

	CIP	Controls
(1) graduated	44%	11%
(2) in-school	23%	12%
(3) drop-out	31%	77%
	N = 180	N = 64

School Retention: The difference in favorable outcomes (graduation, in-school, or transfer) and unfavorable outcomes (dropout) between CIP and controls is statistically reliable ($X^2 = 39.37$, df 1, $p < .001$). Reducing a 77% drop-out rate to 31% was regarded by the high school principals and LEA superintendents who reviewed the findings as meaningful.

TABLE II: MEAN RAW SCORES FOR CIP AND CONTROL STUDENTS

Measure	N	Pretest Mean	SD	Post-test Mean	SD	Change	ANCOVA F&P
Reading							
CIP	88	28.5	13.5	30.6	14.8	+2.1	F= 9.5
Control	55	28.4	11.9	24.5	17.6	-3.9	p < .01
Math							
CIP	86	14.4	8.1	16.0	8.2	+1.6	F=11.9
Control	54	13.6	9.1	12.3	9.0	-1.3	p < .01
Career Resources for Exploration							
CIP	84	256.8	54.0	298.9	58.1	+42.1	F=11.0
Control	69	252.6	42.6	261.3	53.9	= 9.0	p < .01
Career Information and Decision Making Skills							
CIP	88	13.1	4.6	16.4	5.3	+ 3.3	F=10.0
Control	66	12.7	4.5	13.8	4.5	+ 1.1	p < .01
Career Planning Orientation							
CIP	84	96.9	22.1	122.9	21.5	+26.0	F=18.8
Control	64	96.8	16.6	108.2	20.8	+11.4	p < .001
Academic-School Self-Esteem							
CIP	86	4.1	1.7	5.4	2.0	+ 1.3	F= .21
Control	66	3.7	1.6	5.1	1.8	+ 1.4	N.S

Basic Skills Achievement: Table II shows that CIP interns gained more on both reading and mathematics achievement and that the difference is statistically reliable. With regard to meaningfulness, in raw score terms, the gain difference in reading achievement is 6 points (+2.1) - (-3.9), which is about 1/2 of a raw score standard deviation. (The mathematics gain difference is 2.9 points (+1.6) - (1.3), slightly less than 1/2 of a raw score standard deviation.) Regarding grade equivalents, there is about one year grade equivalent difference in both reading and mathematics favoring the Interns.

Career Information, Planning, and Decision-Making Skills: Table II shows that CIP interns gained more on the three sub-scales of the Super Career Maturity Inventory and that the difference is statistically reliable. With regard to educational meaningfulness, the gain difference in Use of Resources (42.1 - 9.0 + 33.1) is more than 1/2 a raw score standard deviation. On Information and Decision-making, the gain difference (3.3 - 1.1 = 2.2) is slightly less than 1/2 a raw score standard deviation. On Planning Orientation, the gain difference (26.0 - 11.4 + 14.6) is more than 1/2 a raw score standard deviation.

When compared with the general population norm groups, the interns' final levels of achievement averaged in the upper middle to very high ranges; the controls' final levels averaged in the lower and lower middle ranges.

The statistical differences are consistent with observations of the ethnographers, as suggested by two abbreviated statements on career views of an intern and a control student (selected as representative from the 24 in-depth case studies).

> Denise: "As long as I could remember, I always wanted to be in a courtroom somewhere, so that's why I said court reporter first. My mind hasn't changed but two of my careers are in conflict and I can't make up my mind at this point. The probation officer would be like a counselor and I like helping people. The court reporter would be in court listening to everything and you have to work under a great deal of stress and you have to really be together upstairs because if you are not--and are not accurate--then I don't think the job is for you. With regard to prestige, I think I feel as though they are about the same, even though other people may not. For me, what's important is if I'm able to do the job, and feel comfortable doing it, and if I'm able to do everything possible to bring out the best in myself and help others."

> Diane: "I've been interested in acting since I was in the sixth grade. It gives you a sense of power, getting up on stage and I'm interested in law because I am interested in going into international politics. I'll make more money going into law than I can make in acting, but I can do both. I'll get my law degree first then go on to acting because I can always have something to fall back on. Everybody wants a lawyer. I'm looking for personal satisfaction. I's hard not to make it as a lawyer if you can pass the bar. And I am going into international law, which is practical and something to fall back on. Or I could work as a court clerk. Court clerks take notes for a trial, carry around top secret messages and such."

Denise was a student at CIP interviewed just prior to graduation. Diane was a member of the control group and at the time she was interviewed, was a twelfth grade student at one of the larger Philadelphia high schools.

School-Academic Self-Esteem: The Academic-School Self-Esteem Scale asks the respondents how they feel about themselves as learners and members of the academic community. The means and variances were uniform between the two groups over the two testings. The change that was found was unexpected inasmuch as both groups gained the same amount. There was only .1 of a point difference between the gains made by the groups. This was an unexpected finding because one of the objectives of the program was to improve academic self-esteem. Almost everyone who applied to the program shared a low level school-academic self-esteem, so there was room for improvement. The dynamics that affected this factor were both more complex and different than had been expected.

Interview and case study data collected as part of the ethnographic evaluation suggest the reasons that students in each group grew approximately the same amount on this dimension. These data indicate that a major factor motivating people to apply to the CIP was the low level of academic self-esteem. Twelve months later, at the time of the posttest, the feelings of low self-esteem in this dimension had diminished and were not as intense. Interns had entered an alternative academic situation where they experienced success and had their self-images as learners lifted. People assigned to the control group by and large were removed from an academic setting and were no longer experiencing the academic failure associated with that setting. Students in the control group, then, most of whom had dropped out of school at the time posttests were administered, indicated in interviews, that they no longer felt stigmatized as "poor students", hence their academic-school self-esteem had increased. Thus the academic-school self-esteem of both groups improved and no difference between the groups was found.

In addition to academic self-esteem, the evaluation of the program also looked at the following self-esteem factors: home-parent, self-esteem; general self-esteem; and peer self-esteem. Prior to administration of the pre-test, it was assumed that applicants would demonstrate ample room to grow on each of the scales. Such was not the case. Analyses of the pre-test data on the aforementioned three scales indicated that the large majority of applicants scored extremely high on these dimensions, which reduced the likelihood of significant growth at time post-test. Applicants felt they had very good relationships with their family and friends and had positive views of their own worth. Post-test socres showed no change from the high initial levels.

Early Follow-Up: Six months after graduating from CIP, interns and controls were equally likely to be employed (33% vs 33%). Interns were more likely to enroll in post-secondary education than controls (43% vs. 15%) and less likely to be unemployed (26% vs 44%). For males these differences were much more pronounced particularly with regard to post-secondary enrollment (29% vs. 0%). The difference between favorable outcomes (employment, or enrolled in post-secondary education) and unfavorable outcomes (unemployed) favoring CIP graduates is statistically reliable ($X^2 = 14.13$, df 1, $p < .001$).

B. Generalizability: Using Participant observer techniques for CIP control schools, the evaluators have concluded that to reproduce the effects reported here one probably would need both the OIC/A guidance and the program as structured. Staff, including the CIP principal, changed during the evaluation and the findings were not affected, so there is some evidence that the effects do not depend entirely on unique leadership or teachers. There is no direct test of the reproducibility of effects in another site. Interns did come from several sending schools in Philadelphia and there was no sending school effect in these findings. Also, the entry level characteristics of students are similar to those reported for minority, low income youth in other studies. Unique characteristics of the students do not appear to be a limiting factor in exportability.

Although intended for use by public and private educators, CIP has been tested only when operated by one of OIC/A's 140 affiliate chapters. There are OIC affiliates in every SMSA, most of which are stated to have good relations with LEA's. Even if claims are limited to the same kind of LEA/OICs/A liaison as existed in Philadelphia, it is estimated that about 80% of Black youth between ages of 16 and 21 could be reached by such an arrangement through the 140 affiliates.

We speculate that the exportable and necessary features include (1) small size, (2) individual attention, (3) determination and caring attitude of OICs/A ethos, (4) the career orientation which is observed by ethographers to influence motivation, (5) the structured, focused academic program, (6) the transition support, and (7) competent, motivated staff and leadership.

C. Attribution of Effects: Use of the control group helps associate changes with participation. Rival hypotheses include however, (a) a once-only "lucky" cohort or the enthusiasm of a start-up effect, (b) artifacts in the mode of analysis, (3) decreased motivation on post-testing for controls, or (d) residuals from sending schools, suggesting the gains demonstrated by CIP

students are a result of programs which they experienced at their sending high schools. To assess the contributions of these sources of variation, analyses were run which showed:

(a) The results reported here for pooled groups are also found if separate analyses are run for the three cohorts. The study period was three years and start-up (or wind-up) enthusiasms or depression did not appear to influence the results reliably. If the p values of the three cohorts are regarded as independent replications in the same site, p is less than .001 for the effects reported.

(b) The data were analyzed in three separate ways since no one approach to adjusting for attrition and lack of all data on all subjects seemed perfect. The findings were the same, regardless of analytic technique. The statistical consultant (Dr. Andrew Porter) had concluded the effects are not artifacts of analytic techniques.

(c) With regard to motivation, it was though this might affect post-test data. The findings were still obtained when a correction for guessing was made, and became even stronger, since CIP interns tended to work more slowly but get items correct if they tried them at all.

(d) With regard to residuals from sending schools, the control students did not show such gains. Blocking was done on a number of variables such as initial levels and indicated the effects of CIP were robust over these variables. It is true there were somewhat more dropouts to begin with in the control group despite the random assignment. This was factored into the analysis with the same results on CIP outcomes. When the proportion of in-school and dropouts were weighted, the effects of CIP on school retention varied by only two percent of the total. Whatever the biases were, they didn't have much impact on the findings and conclusions.

(e) Regarding differential attrition, the interns who dropped out of E and C groups' data collection were not different on pretest measures than interns for whom pre and post data were available. Differential attrition of less able students does not seem to be a plausible explanations for the findings.

What Would Be Disseminated: The program is described in "The Career Intern Program: Interim Report" and in the "Career Intern Program: Final Report" available from NIE. Descriptions re based on three years of ethonographic observation, studies on-site visits, and analysis of written materials. Manuals and guidebooks describing how to install and operate a CIP program have been prepared by OICs/A with the assistance of Research for Better Schools, Inc. OICs/A believes that technical assistance and retraining are required in addition to these manuals to operate a CIP. No claims are made as to which element is most essential; the program was developed and tested as a whole.

VII. Costs for Adoption

TABLE III COST INFORMATION PER LEARNER

	INSTALLATION (Non-recurring Costs)	SUBSEQUENT YEARS (Recurring Cost)
Staff Training	$ 135.66	$ 96.90
Special Facilities	755.81	658.91
Equipment	310.08	69.77
Consumables	174.42	62.02
Other Costs	1,937.98	1,496.12
TOTAL COSTS	$3,313.95	$2,383.72

References

1) Langsdorf, Michael G., et al; The Career Intern Program; Preliminary Results of An Experiment in Career Education (Interim Report, Vols. I, II)

2) Langsdorf, Michael G., et al; The Career Intern Program; An Experiment In Career Education; NIE Papers in Education and Work: Number Seven (Final Report, Vols. I, II)

3) Opportunities Industrialization Center of America and Research For Better Schools: The Career Intern Program Implementation Planning Guide

4) Opportunities Industrialization Centers of America and Research for Better Schools: UCEC; A New Meaning for Education

REFERENCES

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. (Also published as Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

Horst, D. P., Tallamadge, G. K., & Wood, C. T. A practical guide for measuring project impact on student achievement. Washington, D. C.: U. S. Government Printing Office, 1975. (Stock No. 017-080-01460)

Saretsky, G. The OEO P. C. experiment and the John Henry effect. Phi Delta Kappan, 1972, 579-581.

Tallmadge, G. K., and Horst, D. P. A procedural guide for validating achievement gains in educational projects. Washington, D. C.: U. S. Government Printing Office, 1976. (Stock No. 017-080-01516)

SOME SOCIAL FAIRNESS GUIDELINES

Guidelines for non-sexist language in American Psychological Association journals. American Psychological Association, 1200 19th St. NW, Washington, DC.

SWRL instructional product selection kit: booklet on social fairness considerations. SWRL Educational Research and Development, 4665 Lampson Avenue, Los Alamitos, CA 90720

Guidelines for sex-fair vocational education materials. Advisory Council on Women's Educational Programs, 1832 M Street, NW, Washington, DC 20208

Perspectives on school print materials: ethnic, non-sexist, and others. STRIDE, 1855 Folsom Street, San Francisco, CA 94103

Statement on bias-free materials. Association of American Publishers, One Park Avenue, New York, NY 10016

Guidelines for creating positive sexual and racial images in educational materials. Macmillan Publishing Col, Inc., 866 Third Avenue, New York, NY 10022

Guidelines for the development of elementary and secondary instructional materials: the treatment of sex roles. Holt, Rinehart and Winston, 383 Madison Avenue, New York, NY 10017

Undoing sex stereotypes: research and resources for teachers. Capitol Publications, Inc. 2340 Pennsylvania Avenue, NW, Washington, DC 20037

Analyzing children's books with a chicano perspective. Far West Laboratory for Educational Research and Development, 1855 Folsom Street, San Francisco, CA 94103

Content analysis of textbooks for black students, grades 1 to 3. Far West Laboratory for Educational Research and Development, 1855 Folsom Street, San Francisco, CA 94103

Selecting children's books with a black perspective. Far West Laboratory for Educational Research and Development. 1855 Folsom Street, San Francisco, CA 94103

I wonder where the yellow went? Distortions and ommissions of Asian Americans in California Education. ERIC: ED 093 747.

Textbooks and the American Indian. The Indian Historian, 1451 Masonic Avenue, San Francisco, CA 94117

Guidelines for the representation of exceptional persons in educational material. Council for Exceptional Children, 1920 Association Drive, Reston, VA 22091